

I.P. Stepanenko

FUNDAMENTALS OF MICROELECTRONICS

MIR PUBLISHERS
Moscow



FUNDAMENTALS OF MICROELECTRONICS



И. П. Степаненко

ОСНОВЫ МИКРОЭЛЕКТРОНИКИ

**Издательство «Советское радио»
Москва**

I.P. Stepanenko

FUNDAMENTALS OF MICROELECTRONICS

Translated from the Russian
by
P.S. Ivanov

**Mir Publishers
Moscow**

First published 1982
Revised from the 1980 Russian edition

The Greek Alphabet

A α	Alpha	I ι	Iota	P ρ	Rho
B β	Beta	K κ	Kappa	Σ σ	Sigma
Γ γ	Gamma	Λ λ	Lambda	T τ	Tau
Δ δ	Delta	M μ	Mu	Υ υ	Upsilon
E ε	Epsilon	N ν	Nu	Φ φ	Phi
Z ζ	Zeta	Ξ ξ	Xi	X χ	Chi
H η	Eta	O ο	Omicron	Ψ ψ	Psi
Θ θ	Theta	Π π	Pi	Ω ω	Omega

На английском языке

(d) Издательство «Советское радио», 1980
(d) English translation, Mir publishers, 1982

CONTENTS

Preface	7
Chapter 1. Basic Concepts of Microelectronics	9
1.1. General	9
1.2. Integrated Circuits	12
1.3. Main Features of Integrated Circuits	18
1.4. Conclusion	20
Chapter 2. Semiconductors	22
2.1. General	22
2.2. Structure of Semiconductors	22
2.3. Charge Carriers	28
2.4. Energy Levels and Bands	30
2.5. Electric Conductivity	43
2.6. Carrier Recombination	51
2.7. Field Effect	58
2.8. Behavior of Carriers in Semiconductors	65
Chapter 3. Semiconductor Junctions and Contacts	76
3.1. General	76
3.2. Electron-Hole Junctions	76
3.3. Transient Behavior of pn Junctions	97
3.4. Semiconductor-Metal Contacts	106
3.5. Semiconductor-Insulator Interface	112
Chapter 4. Bipolar Transistors	116
4.1. General	116
4.2. Transistor Action	116
4.3. Carrier Distribution	120
4.4. Current Gain	127
4.5. Static Characteristics	133
4.6. Small-Signal Circuit Models and Parameters	139
4.7. Transient and Frequency Characteristics	142
Chapter 5. Unipolar Transistors	150
5.1. General	150
5.2. MOS Field Effect Transistors	151
5.3. Junction Field Effect Transistors	166
Chapter 6. Basics of Microelectronic Technology	173
6.1. General	173
6.2. Preliminary Operations	173
6.3. Epitaxy	174
6.4. Thermal Oxidation	176
6.5. Doping	178
6.6. Etching	185
6.7. Masking	188
6.8. Thin Film Deposition	194
6.9. Metallization	199

6.10. Assembling	201
6.11. Thin-Film Hybrid IC Technology	204
6.12. Thick-Film Hybrid IC Technology	209
Chapter 7. Integrated Elements	212
7.1. General	212
7.2. Isolation of Circuit Elements	213
7.3. <i>NPN</i> Transistors	221
7.4. Varieties of <i>NPN</i> Transistors	229
7.5. <i>PNP</i> Transistors	234
7.6. Integrated Diodes	237
7.7. Junction Field-Effect Transistors	239
7.8. MOS Transistors	241
7.9. Semiconductor Resistors	249
7.10. Semiconductor Capacitors	254
7.11. Film Integrated Elements	259
Chapter 8. Basics of Digital Circuit Engineering	264
8.1. General	264
8.2. Classification of Electronic Circuits	264
8.3. Static Operation of a Simple Bipolar Switch	266
8.4. Transients in a Simple Bipolar Switch	274
8.5. Schottky-Barrier Transistor Switch	282
8.6. Current Switch	284
8.7. MOS Transistor Switches	288
8.8. Noise Immunity of Switches	297
8.9. Bistable Units and Flip-Flops	301
8.10. Schmitt Trigger	308
Chapter 9. Basics of Analog Circuit Engineering	311
9.1. General	311
9.2. Composite Transistors	313
9.3. Statics of a Simple Amplifier	314
9.4. Transients in a Simple Amplifier	321
9.5. Simple MOSFET Amplifiers	325
9.6. Differential Amplifiers	330
9.7. Emitter Followers	344
9.8. Cascode Amplifier	354
9.9. Output Stages	355
9.10. Voltage Regulators	361
9.11. Current Regulators	366
Chapter 10. Integrated Circuits	374
10.1. General	374
10.2. Bipolar Logic Elements	374
10.3. Integrated Injection Logic	385
10.4. MOS Logic	390
10.5. Logic Element Parameters	396
10.6. IC Flip-Flops	400
10.7. Memories	407
10.8. Large Scale Integration	413
10.9. Charge-Coupled Devices	420
10.10. Operational Amplifiers	427
10.11. Testing of Integrated Circuits	440
10.12. Reliability of Integrated Circuits	442
10.13. Conclusion	447
References	449
Index	450

PREFACE

Microelectronics as a logical extension of electronics is distinguished for organic unity of its physical, technological, and circuitry aspects. For this reason, it is hardly possible to expect creative development of integrated circuits relying, for example, only on the effort of "classical" engineers concerned with designing individual semiconductor devices but not having the background in microcircuitry. Also, the effort of "classical" circuit design engineers alone is not enough for further advancement. An engineer engaged in microelectronics must equally well know the basics of microelectronic physics, technology, and microcircuitry. Only with this background can he specialize in any branch of microelectronics.

For the last ten years a few excellent textbooks have been issued for students specializing in microelectronics. Unfortunately, they deal either with physical aspects or only physical-technological and design aspects and cover but to a much lesser extent the elements of integrated circuits and hardly touch on circuit engineering.

This textbook is an attempt to characterize more or less fully all the constituent parts of microelectronics. Particular attention is given to the aspects which have not been treated in detail in other books on the subject. These are transistor physics, transistor microcircuitry, and integrated circuitry in general. The limited space of the book did not allow for the equally detailed description of the design, manufacturing, metric, and some other problems. But this does not seem to be a severe limitation since the reader can obtain additional information from the available books.

So, the intention was to organize the book on the basis of a system approach: to interconnect the sections so that they all in combination form the foundation for further specialization of students. As for engineers who have different backgrounds, but must acquire a thorough knowledge of microelectronics, they should focus on sections dealing with physics, technology, and circuitry.

Being guided by the system approach, the author has included into the text all the necessary sections of microelectronics, irrespective of the fact that some of them are the subjects of special courses, such as "Semiconductor physics", "Semiconductor devices", "Pulse circuits", "Amplifiers", and others. Such a viewpoint is shared by many authors. But it would be hardly justifiable to include the information just mechanically even in a highly contracted form. The

material was not only thoroughly selected, but revised in accordance with the concrete problems of microelectronics and with due regard for internal interconnection of the sections.

For example, Ch. 2 does not include such important aspects as elements of quantum-mechanical theory of solids, the Gunn effect, Josephson effect, size effects in thin films, and others. Among many types of transistors, Ch. 4 and Ch. 5 consider only low-power transistors which form the basis of modern integrated circuits. The number of transistor circuits dealt with in the book is sharply cut down to place primary emphasis on the concrete configurations employed in microelectronics. The traditional sequence in which two basic classes of circuits are described is changed: Ch. 8 deals with pulse circuits and Ch. 9 with amplifying circuits.

The choice of the right material and the development of the technique for its comprehensive presentation in Ch. 10 posed an equally complex task. To avoid unjustifiable intricacies, it was found necessary to disregard in many cases schematic diagrams and focus on block diagrams. Such an approach is in agreement with the techniques of designing modern LSI circuits.

The author hopes that the course "Fundamentals of microelectronics" built in accordance with the structure of this book will become as traditional as the courses "Semiconductor devices" or "Basic theory of electrical engineering". This would undoubtedly aid both in more adequate training of students of the same specialty and also in promoting understanding between engineers of related specialities, who inevitably come in contact in designing complex microelectronic circuits.

The author wishes to express his deep gratitude to professor C.Ya. Shats, professor V.N. Dulin, and associate professor Yu.E. Naumov for valuable suggestions and ideas and also to the members of the Chair of Microelectronics at the Moscow Engineering Physics Institute for help in drawing up the manuscript.

1.1. General

The role of electronics in the development of modern science and technology can scarcely be overestimated. Electronics is by right considered a catalyst of scientific and technical progress. Indeed electronics is responsible for the achievements in exploration of outer space and deep sea, advancements in atomic power and computer engineering, radio broadcasting and television, automatization of production processes, and studies on living organisms. Microelectronics is the next stage of development of electronics and one of its basic branches that in principle offers new approaches to the solution of problems confronting scientists and engineers nowadays.

Electronics is the field of science and engineering that deals with the research, development of electronic devices and their utilization.

Microelectronics is a branch of electronics concerned with the research, development, and utilization of qualitatively new electronic devices known as integrated microcircuits.

An *integrated microcircuit* (or simply an integrated circuit) is a combination of a few interconnected circuit elements such as transistors, diodes, capacitors, and resistors produced in a single manufacturing process (simultaneously) on one and the same bearing structure, called the *substrate*, and intended to perform a definite function involved in converting information.

If an integrated circuit includes only one type of components, such as only diodes or resistors, it is said to be an *assembly* or *set* of components.

The term integrated circuit (IC) implies a union of components which form a structurally integral device designed to perform more complex functions than those assigned to isolated components.

The inseparably associated and electrically interconnected components that make up an IC are called *integrated (circuit) elements*. They show certain features which distinguish them from conventional transistors or resistors fabricated as structurally individual elements and interconnected by soldering to form a circuit. Such structurally isolated parts typical of the "premicroelectronic age" are known as *discrete components*, and electronic units and blocks based on these components as *discrete circuits*.

What underlies the development of electronics is an ever growing demand for electronic equipments capable of performing more and more complex functions. At a certain stage of technical progress old means such as vacuum tubes and discrete transistors, or what is called old *hardware*, become unsuitable for the solution of new problems. The basic factors which necessitate the replacement of old hardware are the reliability, size, mass, cost, and power requirements. A simple and tentative calculation can reveal the causes of a gradual swing from transistor engineering to microelectronics.

Let it be required to make a compact electronic system consisting of 10^8 components. If we attempt to do the task using discrete components each of which has on the average a power of 15 mW, measures 1 cm^3 with interconnections, weighs 1 g, costs 50 kopecks, and shows a failure probability of 10^{-5} h^{-1} , the device so constructed will dissipate 1.5 MW, measure 100 m^3 , weigh 100 t, and will be worth 50 million rubles excluding the labor cost.

As is clear, the system proves far from being compact and totally unsuitable for, say, airborne applications. It consumes a huge amount of energy which, by the way, it is impossible to dissipate within the system of such dimensions. It is easy to calculate that its assembly will require not less than 10 man-years. The cost of the device is rather high and its production, even in a small batch, may turn out to be too difficult or too unprofitable for the national economy.

But the main conclusion drawn from the example lies in the fact that the mean rate of failures ($10^{-5} \times 10^8$) is equal to 10^3 h^{-1} , that is, about 1 failure in 3 seconds, which of course points to the unserviceability of the device.

It is likewise possible to illustrate the causes of transition from vacuum-tube to transistor technology.

The example considered above shows that we cannot handle the task by means of discrete transistor engineering. To solve the problem, we have to rely on *qualitatively new* hardware which would ensure a decrease in the failure probability, cost, and size by a few orders of magnitude. It is integrated circuits that constitute such hardware.

It should be pointed out that microelectronics whose underlying principle is the integration of components *has arisen from discrete transistor engineering* and adopted its progressive methods and means. Indeed, an integrated circuit owes its appearance to the *batch processing technique* and *planar technology* both put to industrial use in discrete transistor engineering in the late fifties. The state of the art at the time was quite ripe for the onset of microelectronics.

The idea of technological integration of components on a single substrate naturally stemmed from the batch processing and fabrication technique. In principle the technique consists in the following. A large number of transistors are manufactured *simultaneously*

(and arranged in a regular fashion) on a wafer of silicon or germanium 25 to 40 mm in diameter or above (Fig. 1.1*a*). The wafer is then scribed vertically and horizontally and broken into hundreds of individual chips with a transistor in each chip (Fig. 1.1*b*). Next the chips are housed in cases with external leads (Fig. 1.1*c*) and delivered

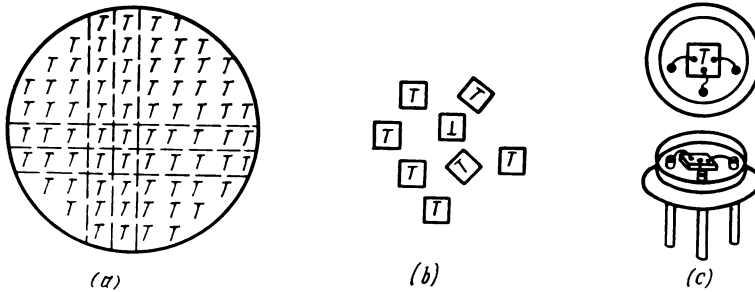


Fig. 1.1. Batch technique for manufacture of transistors

(a) silicon wafer with transistors; (b) individual transistor chips; (c) encased transistor chip

to the development engineer who has to perform the reverse operation, namely, interconnect the transistors (and other components) by soldering to produce the desired functional unit such as an amplifier, memory cell, etc.

The principle of integration of circuit elements lies in the following. Instead of the manufacture of individual transistors, a great

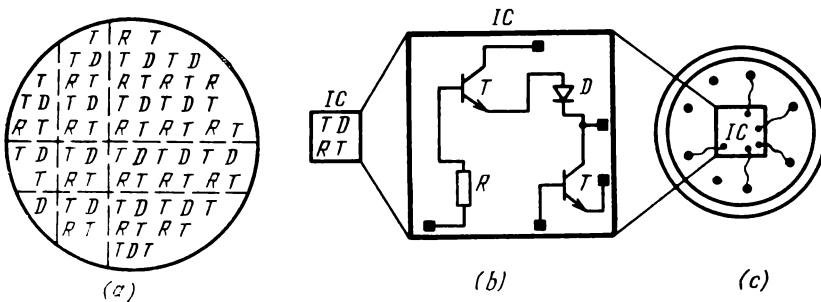


Fig. 1.2. Batch technique for manufacture of ICs

(a) silicon wafer with sets of elements; (b) interconnections of elements; (c) encased IC chip

number of "sets" are produced simultaneously on a wafer. Each set contains all the components such as transistors, diodes, and resistors which make up an appropriate functional block (1.2*a*). The method dispenses with thin wires and soldered leads for interconnection of components. Instead, the components are interconnected with short

fine metallic stripes deposited on the wafer surface. Thus each set of components is a ready integrated circuit (Fig. 1.2b). All ICs are regularly distributed on the wafer surface. It now remains to cut the slice into individual chips and encapsulate each (Fig. 1.2c). The development engineer receives the finished functional unit as a *structurally integral* electronic device.

To ensure interconnection of elements with fine metallic stripes, the leads of all electrodes should lie in one plane, that is, on one and the same surface of the slice. *Planar* technology (see Sec. 6.9), which was one of the latest developments in transistor engineering, provides for such an arrangement of leads. Microelectronics has naturally adapted the planar process along with the batch processing technique for the fabrication of ICs.

Thus in the development and fabrication of IC-based equipments, microelectronic technology does away with numerous soldered connections, which are the basic source of failures, sharply decreases the size and mass of devices since each integrated element is free from the case and lead-out wires, and drastically cuts down the cost of finished products because the process of fabrication obviates the necessity for many packaging and mounting operations. As will be shown later in more detail, these factors combined with improved reliability are a great asset of ICs.

1.2. Integrated Circuits

In the course of development of microelectronics, starting in 1960, the range of IC was changing uninterruptedly. Various types of IC were sometimes regarded as alternative, or incompatible with all other types. At present each of the basic types has its own, relatively stable, place in microelectronics. In the above discourse devoted to the general idea of integration of circuit components, by the basic type of IC we have meant the semiconductor type.

1.2.1. Classification of ICs. By the method of fabrication and the resultant structure, we can distinguish two principally different types of IC, semiconductor and film types.

A *semiconductor IC* is a microcircuit whose elements are fabricated within the surface layer of a *semiconductor substrate* (Fig. 1.3). These ICs constitute the basis of modern microelectronics.

A *film IC* is a microcircuit whose elements are various films formed by deposition on an *insulating substrate* (Fig. 1.4). By the method of film deposition and the thickness of deposited films, film ICs are divided into *thin-film* integrated circuits with the thickness of films up to 1 or 2 μm and *thick-film* integrated circuits with a film thickness ranging from 10 to 20 μm and above. Since none of the combinations of deposited films can so far provide for **active** elements of the transistor type, film ICs contain only **passive elements** such as resistors and

capacitors, for which reason purely film integrated circuits perform limited functions. To obviate this limitation, a film IC is supplemented with active **discrete** components arranged on the same substrate and connected to film circuit elements (Fig. 1.5). Such a composite, film-discrete circuit is called hybrid.

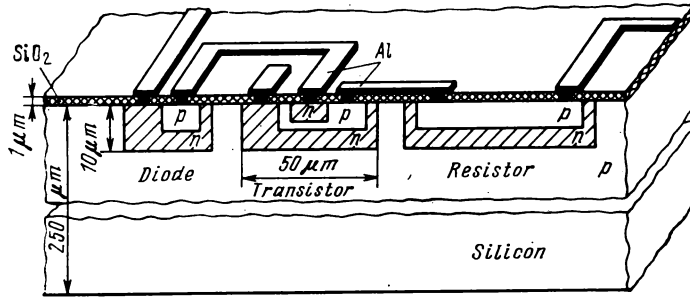


Fig. 1.3. Structure of semiconductor IC

A *hybrid IC* (HIC) is a microcircuit which represents a combination of film passive elements and discrete active elements disposed on a common insulating substrate. Discrete components which form part of a hybrid IC are called *inserted*, or *add-on* elements to stress the fact that they **have** nothing to do with the technological cycle

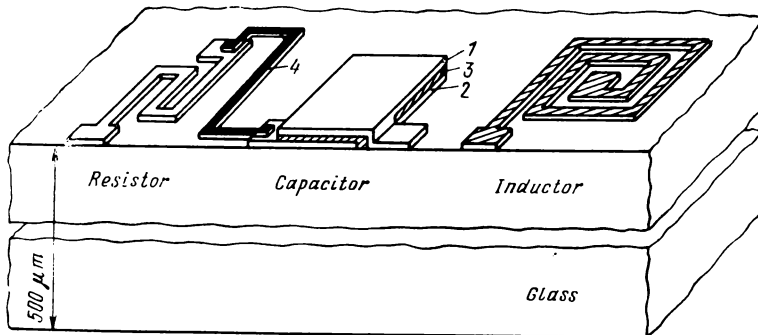


Fig. 1.4. Structure of film IC

1—upper plate; 2—lower plate; 3—dielectric; 4—connection strip

involved in the production of film circuit elements proper. Along with diodes and transistors, a hybrid IC may also contain semiconductor ICs as active components designed to serve more complex functions.

One more type of composite IC which comprises semiconductor elements and film passive elements is a compatible hybrid IC.

A *compatible hybrid IC* is a microcircuit in which active elements are formed within the surface layer of a semiconductor chip (as in a semiconductor IC) and passive elements are deposited as films on the preliminarily insulated surface of the same chip (as in a film IC).

Compatible HICs offer advantages where there is a demand for high values and high stability of resistances and capacitances; these requirements are easier to meet with the aid of film elements than with semiconductor elements.

In all the types of IC, interconnections of elements are accomplished with fine metallic stripes evaporated or sputtered in definite regions

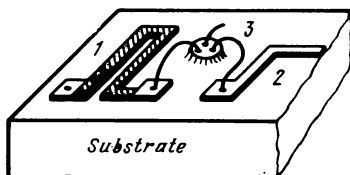


Fig. 1.5. Structure of hybrid IC
1—resistor; 2—metallized strip; 3—discrete transistor chip

on to the surface of a substrate to connect the elements together. The process of deposition of these stripes is known as *metallization*, and the interconnecting metallic pattern as the *connection layout*.

1.2.2. Semiconductor ICs. There are two classes of semiconductor ICs commercially available today, *bipolar ICs* and *metal-insulator-semiconductor (MIS) ICs*. Since in most cases the insulator is a silicon oxide, in the further discussion we shall use the term MOSICs instead of MISICs. A combination of bipolar and MOS transistors within a single chip represents a particular case.

Integrated-circuit technology for both classes of ICs relies on alternate doping of a semiconductor (silicon) slice with donor and acceptor impurities to form within the body of the slice near its surface thin layers of different conductivity types and *pn* junctions at the boundaries of layers. Individual layers serve as resistors, and *pn* junctions as constituent elements of diode and transistor structures.

Doping of a slice with impurities has to be carried out **locally**, that is, in separate diffusion regions spaced at considerable intervals ranging from 10 to 100 μm . The method of local doping makes use of special *masks* with windows for the diffusion of impurity atoms into the desired regions on the slice. In the manufacture of semiconductor ICs, it is usual to employ a film of silicon dioxide, SiO_2 , as a mask that covers the surface of a silicon wafer. A suitable combination of photographic and etching techniques (see Sec. 6.7) provides the required combination of *windows* or, what is called, the *pattern* (Fig. 1.6).

Let us now define in brief the constituent elements of monolithic integrated circuits of both classes.

The basic element of a bipolar IC is an *npn* transistor; the entire production cycle is tailored to manufacture just this element; all other elements must be made, wherever possible, at the same time as the transistor to do away with additional production stages. Resistors are formed at the same time as the base region and thus grown to the

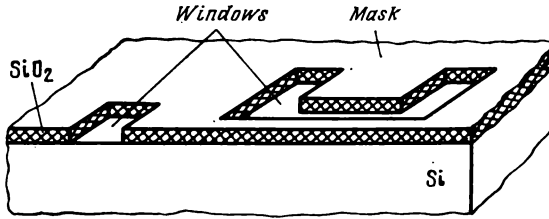


Fig. 1.6. Oxide mask with windows for local doping

same depth as the base. Reverse biased *pn* junctions serve as capacitors, in which *n*-type layers correspond to the collector region of the *npn* transistor, and *p*-type layers to the base region.

The basic element of a MOSIC is a MOS transistor with an induced channel (see Fig. 5.2). MOS transistors connected in a one-port network act as resistors and MOS structures play the role of capacitors; in these capacitors, the dielectric layer is formed simultaneously with the gate dielectric and the semiconductor "plate" simultaneously with the source and drain regions.

The elements of a bipolar IC should be suitably isolated to prevent them from interaction over the bulk of the chip. Isolation techniques for bipolar devices are considered in Sec. 7.2. The elements of MOSICs do not require isolation because no interaction exists between adjacent MOS transistors (see Fig. 7.2) and they can be spaced a minimum distance apart, this being one of the basic advantages of MOSICs over bipolar ICs.

The characteristic feature of a semiconductor IC is the absence of inductors and transformers among its circuit elements. The attempts to use any physical phenomena in a solid body as an equivalent to electromagnetic induction have failed so far. In the development of ICs, therefore, the efforts are made to implement the desired function without resorting to inductances. In most cases this approach gives fruitful results. Where the circuit design inevitably requires the use of an inductor or transformer, suitable attachment of a discrete element to the substrate helps solve the problem.

The chips for modern semiconductor ICs measure from 1.5 by 1.5 mm² to 6 by 6 mm². A larger area of the chip permits us to lay

out a more complex IC containing more components. It is possible to dispose a greater number of elements on the same area of the chip by decreasing the dimensions of elements and spacings between them.

It is customary to define the functional complexity of ICs by the *scale of integration*, that is, by the number of elements (most often transistors) within a chip. In 1980-1981, the maximum scale of integration reached 10^5 elements in a chip. An increase in the scale of integration (and thus in the complexity of functions performed by ICs) is one of the main trends in microelectronics.

The scale of integration is sometimes quantitatively expressed through a factor $k = \log N$, where N is the scale of integration. If $k \leq 1$ (that is, $N \leq 10$), the circuit is said to be a *small-scale* IC; if $1 < k \leq 2$, this is a *medium-scale* integrated circuit or MSI; if $2 > k \leq 3$, this is a *large-scale* integrated circuit or LSI; and if $k > 3$ (that is, $N > 1\,000$), the circuit is called a *very large-scale* integrated circuit or VLSI.

Along with the scale of integration, in current use is one more index, called the *packing density*, which is the number of elements (commonly transistors) contained in a unit area of the chip. The packing density, which largely characterizes the technological level, at present reaches 1000 to 2000 elements per mm^2 .

1.2.3. Hybrid ICs. As mentioned above, film and hybrid ICs are divided into thick-film and thin-film ICs depending on the technology of production.

Thick-film HICs are fabricated in a simple way, even in a primitive way as it may appear at first glance. Layers of various inks (*pastes*) are deposited on an insulating substrate of a rather large area, a few square centimeters. A distinctive feature of the thick-film technique is the possibility of obtaining *at once* the film of a desired thickness. Conductive pastes provide interconnections, capacitor plates, and contact pads for connection of elements to case leads, or pins. Resistive pastes provide resistors, and dielectric pastes ensure insulation between capacitor plates and protection in general of the surface of the ready HIC. Every layer of paste must be of a certain configuration, or pattern. In the deposition of each layer of film, the paste is applied to the substrate through the windows of a *stencil screen*, or mask, which defines the pattern for the desired film (Fig. 1.7). After the passive circuitry on the substrate is completed, the next step follows which involves bonding of discrete components to the reserved areas or to the protective dielectric layer and connection of their leads to *contact (bonding) pads* formed in conductive layers. Thick-film HIC technology and the parameters of the hybrid circuit elements are considered in more detail in Sec. 6.11 and Sec. 7.11 respectively.

The above short description allows us to single out the following features of the thick-film hybrid approach:

1. The “mechanical” deposition process does not permit fabrication of films less than 10 to 20 μm in thickness (typical thicknesses being 50 to 100 μm), hence the names *thick-film* technology and *thick-film* hybrid circuit.

2. Simple technology¹ makes hybrid circuits more available and cheaper.

3. The “mechanical” method of film deposition cannot ensure sufficiently small tolerances on the ratings of resistors and capacitors, that is, cannot produce precision elements.

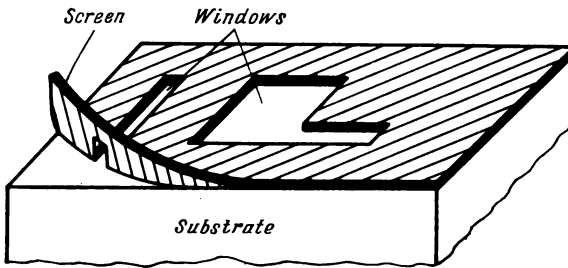


Fig. 1.7. Stencil screen for local deposition of ink

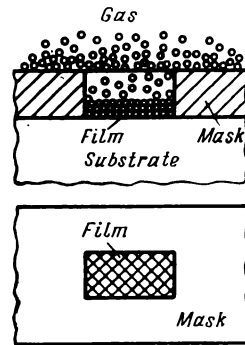


Fig. 1.8. Layer-by-layer growth of thin film

Thin-film HICs necessitate more complex technology than thick-film hybrids. Besides, the thin-film deposition process uses specific tooling and equipment, commonly rather expensive, so thin-film hybrids are costlier than thick-film HICs.

Classical thin-film technology involves deposition of films on a substrate from a *gaseous* phase. In distinction to the deposition of thick films, the growth of thin films to the final thickness proceeds gradually, one monomolecular layer over the other (Fig. 1.8). After deposition of the film, the growth of the next film follows, now from the gaseous phase of a different chemical composition to impart the film the desired electrical and physical properties. This procedure enables the consecutive growth of conductive, resistive, and dielectric layers. A metal mask placed over the substrate (as in the thick-film deposition process) or a mask grown on the surface (similar to an

¹ In reality this technology is of course not so simple as it may appear from the above description. The problem that faces engineers is the search for the right formulation of pastes noted for very complex compositions since they must meet numerous and often conflicting requirements.

oxide layer in the fabrication of semiconductor ICs, as shown in Fig. 1.6) defines the pattern for each film layer.

In order that atoms or molecules of a vapor might move freely from the source to the substrate, the process of film deposition should take place in a confined space with a vacuum of the desired degree.

As with thick-film HICs, discrete elements of thin-film HICs are mounted on to the finished passive circuitry and connected to the proper contact pads. A more detailed description of thin-film HIC technology is given in Sec. 6.10, and the parameters of the thin-film hybrid elements are considered in Sec. 7.11.

The above description permits us to point out the following features of the thin-film hybrid approach:

1. Since films grow at a comparatively low rate, the deposition of films over $1\text{ }\mu\text{m}$ in thickness takes much time. Besides, the deposited films over 1 or $2\text{ }\mu\text{m}$ thick easily peel off. The typical thickness of thin-film HICs does not exceed $0.5\text{--}1\text{ }\mu\text{m}$, hence the terms *thin-film* technology and *thin-film* HIC.

2. A low rate of film growth allows a rather easy control of film thicknesses, ensuring close tolerances on the values of resistors and capacitors and thus a high precision of these elements.

The scale of integration of HICs cannot be estimated in the same way as for semiconductor ICs since there is no film active component here that would serve as a "point of reference". All the same there is a term *large-scale* HIC (LSHIC), which means that the discrete components attached to this circuit are not only transistors but complete semiconductor ICs, and so the large-scale HIC can perform a much more complex function than an individual IC or even an LSI.

1.3. Main Features of Integrated Circuits

An integrated circuit belongs to the family of *electronic devices*; like a vacuum tube or transistor, this device represents a constructional unit, performs a definite function, and must satisfy certain test and performance requirements. But in comparison to, say, a diode or transistor, the IC is a qualitatively new type of device.

The first basic feature of an IC as an electronic device is that the integrated circuit can perform a rather complex function, whereas elementary electronic devices can perform a similar function only in combination with other components. For instance, an individual transistor is unable to amplify a voltage signal or store information. For this, it is necessary to *assemble* an appropriate circuit from a few transistors, resistors, and other components. In microelectronics, however, a single device—an integrated circuit—can act as a voltage amplifier or storage device.

The second important feature of an IC is that an increased functional complexity of this device does not affect any of the basic factors such

as reliability and cost. The contrary is true for elementary discrete devices. Moreover, ICs offer reliability and cost advantages. To illustrate this feature, consider an example of semiconductor ICs.

Since small-scale and medium-scale ICs are comparable with discrete transistors in size and mass, it can be assumed that in a first approximation the gain in size and mass on converting from discrete to integrated circuits depends on the scale of integration and thus may run to hundreds and thousands. But it is difficult to estimate accurately this gain by theoretical calculations because ICs have other standard sizes of enclosures and a larger number of leads than discrete components.

The operating reliability of a semiconductor device included in a functional unit is largely determined by the number of soldered and (to a lesser degree) welded joints. Since the interconnections of integrated elements are made by metallization (that is, without soldering and welding), the IC certainly shows better reliability than the discrete circuit that serves the same function. As the scale of integration grows, the reliability improves.

Because all the elements of an IC are fabricated in a common technological cycle, the number of operations involved in the production of these elements does not greatly exceed the number of operations required to manufacture an individual transistor. For this reason the cost of an IC, other conditions being the same, compares to the cost of a single transistor. So, depending on the scale of integration (or, more precisely, on the packing density), the cost of an integrated circuit element may be as small as hundredths of the cost of a similar discrete component. The same relation exists between the cost of an IC and the cost of a similar circuit based on discrete components.

The third feature of an IC lies in the preference of active elements over passive ones—a feature diametrically opposite to that inherent in discrete transistor techniques. In a discrete circuit the active components, particularly transistors, are most expensive, for which reason the circuit optimization, considering that all other conditions are equal, consists in a decrease in the number of active elements. Not so with an IC: it is the chip that defines the cost rather than an element. Thus it is advantageous to build into a chip as many small-area elements as possible. Active elements such as transistors and diodes occupy the smallest areas, and passive elements the largest. So the optimal integrated circuit is an IC which contains a minimum number of resistors and, particularly, capacitors, both having the lowest nominal values.

The fourth feature of an IC is associated with the fact that adjacent elements are spaced merely 50 to 100 μm apart. At such intervals, the material is not very likely to show differences in electrical and physical properties, and so a considerable difference between the parameters of adjacent elements is improbable. In other words, *the*

parameters of adjacent elements are mutually related, or correlated. This correlation holds with temperature changes too: the temperature coefficients of parameters are practically identical for adjacent elements. The correlation of the parameters of adjacent elements is used in designing some ICs to reduce the effect of spread in values and the effect of temperature changes.

Hybrid ICs also represent a particular type of electronic device. But hybrids incorporate isolated components, so they are less specific than semiconductor ICs. Nevertheless both types have many common features.

The basic feature specific to any type of IC is its *functional complexity* which leads to *qualitative* changes in the structure of electronic equipment. As compared with a semiconductor IC, a HIC may feature either high values of resistors and capacitors, unattainable in semiconductor ICs, or a high precision of resistors, or, last, an increased functional possibilities typical for a large-scale HIC.

The gain in reliability, cost, size, and mass of HICs is due not only to the passive film network but also to the use of unencapsulated active components and to a decreased number of welded joints and assembly operations.

One of the important and distinctive features of HICs lies in the possibility of *correction* (trimming) of resistors to the specified tolerances before circuit completion and encapsulation. This considerably reduces the spread of resistances and enables the fabrication of precision resistors essential for measuring devices and computers. As a whole, a HIC may be considered to be quite a versatile, cheap, and readily producible type of IC well adapted to solving special, particular problems.

1.4. Conclusion

The first stages in the development of microelectronics were primarily characterized by the progress in integrated-circuit technology. The common trend at the time was toward the perfection of isolation methods, LSI technology, and mounting and bonding methods. As regards circuit engineering, in those early stages of microelectronics development, the trend was to adopt the circuit layout techniques from discrete transistor electronics.

But it soon became apparent that circuit designs must be consistent with qualitatively new techniques of IC fabrication. The circuits, regarded as typical in discrete transistor techniques, proved far from being suitable for use in microelectronics. On the other hand, many discrete circuits, considered as being "exotic" and found but limited applications, turned out to be quite appropriate and even optimal for microelectronic devices. That is why integrated-circuit

configurations do not coincide with those typical of conventional transistor techniques.

During the development of microelectronics there have appeared some specific integrated elements such as multiemitter transistors and charge-coupled devices. There are no analogs of these devices in discrete transistor circuits. Integrated circuits that contain these specific elements are not amenable even to modeling with discrete components.

The above discussion points to the fact that microelectronics as a field of science and technology is in no way limited only to IC fabrication. It combines three equivalent aspects: *physics*, *technology*, and *circuit engineering*. The knowledge of these three microelectronic aspects helps the development engineer to estimate judiciously both new variants of the hardware or circuit designs from the viewpoint of their realization and new variants of production processes from the viewpoint of their applicability for the fabrication of certain elements and circuits.

These three aspects are a subject dealt with in subsequent chapters. The last chapter describes some types of IC and trends in microelectronics.

2.1. General

By convention, semiconductors are considered to be substances having a resistivity midway between that of insulators and metals. Semiconductors thus span the resistivity range from 10^{-3} to $10^9 \Omega \text{ cm}$, metals from $10^{-4} \Omega \text{ cm}$ and below, and dielectrics from $10^{10} \Omega \text{ cm}$ and above. Of course, such a purely quantitative classification is rather conditional, particularly as regards semiconductors and dielectrics, since in principle there is no essential difference between these two classes of substances. The differences between semiconductors and metals are more substantial and manifold; they differ in many other parameters apart from resistivity.

Since *semiconductor* devices (primarily transistors) form the basis of modern microelectronics, the physics of semiconductors deserves consideration. For this reason, this Chapter deals entirely with semiconductors; it does not cover metals and dielectrics in a particular way, but certainly considers their basic features to distinguish these materials from semiconductors.

Among semiconducting materials, *silicon* has proved most suitable for the manufacture of integrated circuits. This material rapidly ousted germanium (which had played a historical role in the growth of transistor electronics) and up to now it has not met any worthy rival. In all examples and illustrations, therefore, we shall use electrical and physical parameters of silicon and also point out those properties and features of this material which have enabled it to hold the leading place in microelectronics.

2.2. Structure of Semiconductors

The area of modern ICs is in the order of 2 to 30 mm², the area of circuit elements ranges from 10^{-2} to 10^{-3} mm², and the linear dimensions of individual electrodes are as small as 1 μm . It is obvious that within such areas and spacings a semiconductor wafer must be sufficiently homogeneous and have controlled properties. If a semiconductor yet shows defects and inhomogeneities, these should be local and their number as small as possible in order to limit the number of rejects only to the ICs located within the defective regions. This explains why the problem of obtaining homogeneous, defect-free semiconductor crystals deserves so much attention.

2.2.1. Crystal lattice. Semiconductors are as a rule solid bodies—*single crystals*—displaying a regular crystal structure, whose crystal lattice consists of a great number of periodically repeating adjacent *unit cells* of a certain variety of shapes and sizes. For the simplest cubic lattice (Si, Ge, NaCl, and others), the edge of its unit cell—the cube—is the *lattice constant* a , equal to 0.4 to 0.6 nm. The *diamond-type cubic lattice* (Si, Ge) consists of tetrahedrons (Fig. 2.1), with the distance between neighbor atoms coming to about 0.25 nm.

Bonds between the atoms in the crystal lattice of silicon and a number of other semiconducting materials are attributed to specific mutual forces which arise from the union of pairs of valence electrons of neighbor atoms. Such a bond, in which either of the atoms remains neutral, is called a *covalent*, or *valence*, bond.

The regularity (periodicity) of a crystal structure is the cause of differences between the physical properties of a crystal in different directions. This directional dependence of physical properties is known as *anisotropy*. To estimate various directions in the crystal and thus readily recognize the type of crystalline substance, it is customary to employ *crystallographic axes* and *planes* normal to these axes. In use is a convenient system of designation of crystal axes and planes by three-digit *Miller indexes*. Thus the indexes put in brackets, such as $[111]$, $[100]$, etc., identify axes, and the indexes enclosed in parentheses, such as (111) , (100) , etc., denote planes or faces.

The derivation of Miller indexes for the simplest cubic lattice is illustrated in Fig. 2.2a. As seen, a plane cuts the axes into intercepts measured in units of the lattice constant: $x = la$; $y = ma$; $z = na$; where l , m , and n are integers. By reducing the reciprocals l^{-1} , m^{-1} , and n^{-1} to the least common denominator, we take the numerators as Miller indexes for the given plane.

Note that every crystal plane exhibits a definite density of atoms per unit area. Thus if we “view” a cubic crystal in the direction normal to the planes (100) , (110) , and (111) , the arrangement of atoms that fall within our view will be as shown in Fig. 2.2b (the atoms on the lattice sites being numbered for clarity). The plane (111) shows the highest density of atoms, and the plane (100) the lowest. In silicon, the plane (111) is the *plane of cleavage*, along which the crystal cracks and splits.

Many properties and parameters of a crystal, such as optical characteristics and rate of etching, vary in different crystal planes.

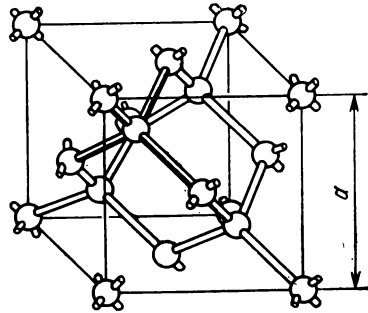


Fig. 2.1. Diamond-type crystal lattice structure

This calls for accurate grinding of a slice for ICs along the chosen crystal plane and control over slice fabrication with the aid of X-ray defraction methods.

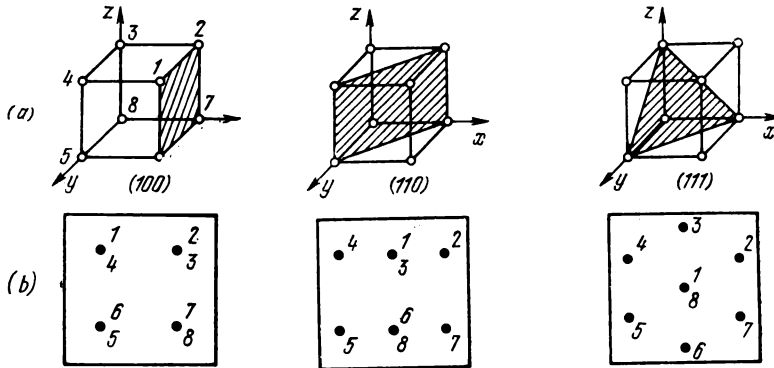


Fig. 2.2. Crystal planes

(a) derivation of Miller indexes; (b) arrangement of atoms in crystal planes

In recent years, *liquid crystals*, that is, anisotropic liquids, have come into use. They have a complex chemical composition and contain organic substances. What makes liquid crystals useful for practical purposes (in various light indicators, such as luminous dials of electronic watches, and in other devices) is the ability to change their optical properties on exposure to an electric field or temperature.

2.2.2. Crystal defects. Whatever the type of crystal, its structure never happens to be perfect; structural imperfections always exist in the bulk of a crystal, let alone the defects on its surface.

Lattice defects may show up as a vacant lattice site (the *Schottky defect*, in which case the atom that has formed a vacancy moves on to the crystal surface) or a combination of the vacant site and interstitial atom (the *Frenkel defect*). These are *point defects* (Fig. 2.3a, b) which inevitably appear in a crystal; the concentration of point defects follows a thermodynamic rule: it sharply grows with temperature. Point defects spread almost uniformly throughout the bulk of a crystal.

Other lattice defects may result from the bombardment of a crystal by heavy nuclear particles, called nucleons. These are *radiation-induced* defects.

Any real semiconductor contains impurities, either harmful, which are difficult to remove in refining, or useful, introduced intentionally to impart the desired properties to the crystal. Every foreign, or stranger, atom is in fact a point defect in the crystal lattice.

Impurity atoms can lie either in interstices (*interstitial impurity*, as 1 in Fig. 2.3c) or on sites themselves as they replace parent atoms (*substitutional impurity*, at 2 in Fig. 2.3c). The latter defect is more widespread.

Dislocations are displacements of lattice **planes**. They can be of the *line* (edge) and *screw* (helical or Burgers) types. The first type

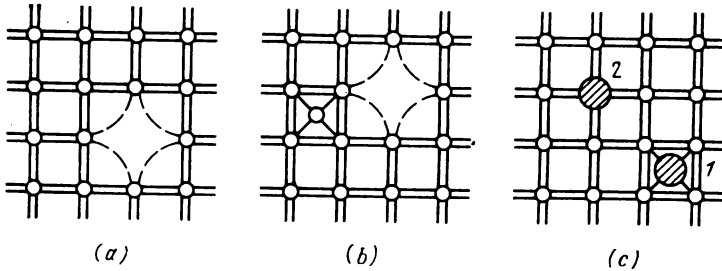


Fig. 2.3. Point defects in the crystal lattice
(a) Schottky defect; (b) Frenkel defect; (c) impurity defects

results from a partial shear (partial slip) of the lattice so that an **incomplete half-plane** of atoms appears (Fig. 2.4a). The second type results from a **complete** shift (over the entire depth) of some portion of the lattice (Fig. 2.4b).

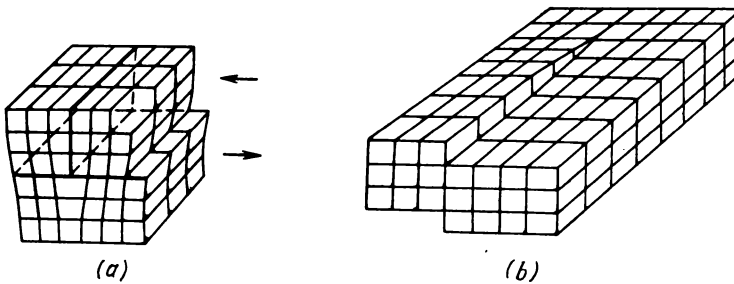


Fig. 2.4. Dislocations in the crystal lattice
(a) line; (b) screw

The density of dislocations is estimated visually (with a microscope) by counting *etch pits*. The formation of pits under the action of an etchant is one of the indications of changes in the properties of a crystal at the site of dislocations. A faster diffusion of impurities into dislocations may serve as another indication of changes in properties. Both examples point to the fact that dislocations may lead to undesirable results in the production process, that is, to uncontrolled profiles (dimensions) of IC elements. *This calls for*

setting strict limits on the number of dislocations in a semiconductor wafer. At present the permissible dislocation density for silicon slices is set at 1 to 10 mm^{-2} , though in the process of treatment it can grow to a certain degree. Over the past few years *dislocation-free* silicon slices have become available. The dislocation density in these slices does not exceed 1 cm^{-2} .

A *polycrystal* may represent a limiting case of **random** dislocations. This crystal consists of a great many of the tightly adjoining monocrystalline grains (microcrystals) with various orientations. Polycrystals lack the periodicity of structure and thus are free from the anisotropy which is intrinsic in single crystals. They can have both a fine-grain and a coarse-grain structure. The size of grain determines many electrical and physical properties, for example, electric conduction. The larger the grains, the smaller the role of boundaries between the grains, and, specifically, the lower the resistivity of a polycrystal. But since the microstructure of polycrystals is practically uncontrollable, the reproducibility of their electrical and physical properties is incomparably poorer than that of single crystals. This is the reason why polycrystals have not become basic materials for most responsible (**active**) integrated elements and play an auxiliary part in microelectronics.

A polycrystal in itself contains a large number of defects, so it behaves rather indifferently to the appearance of new defects induced, for example, by nucleonic radiation. Hence, an enhanced radiation stability of polycrystals.

Along with dislocations, macroscopic defects such as microcracks and pin holes (voids) appear in semiconductor slices. These defects are potential causes of malfunction of ICs.

2.2.3. Crystal surface. Atoms located on the surface of a crystal have a part of their covalent bonds inevitably broken because they lack neighbors on the other side of the interface. The number of broken bonds depends on the crystallographic orientation of the surface. Thus for silicon, one of the four bonds is seen to be incomplete in the plane (111), and two in the plane (100), as shown in Fig. 2.5.

The rupture of covalent bonds entails disturbances in energy equilibrium on the crystal surface. The surface energy comes to equilibrium in a number of ways: (1) through changes in the distance between atoms in the surface layer, that is, changes in the structure of unit cells; (2) through capture—*adsorption*—of foreign atoms from the environment, which fully or partially reconstitute the broken bonds; (3) through the formation of a chemical compound such as an oxide with completely filled bonds on the surface; etc. In any of the cases, *the structure of a thin surface layer*, about a few nanometers thick or even less, *differs from the structure in the crystal bulk*.

In consequence, the electrical and physical properties of a surface layer noticeably differ from the properties of the bulk; this conclusion holds whether a crystal is in a vacuum, atmosphere, or adjacent to a certain solid body. *The surface, or boundary, layer* (it is customary to refer to it simply as the surface, or boundary) *should thus be regarded as a specific region of crystal*. This region plays an important part in microelectronics since the elements of planar ICs are located

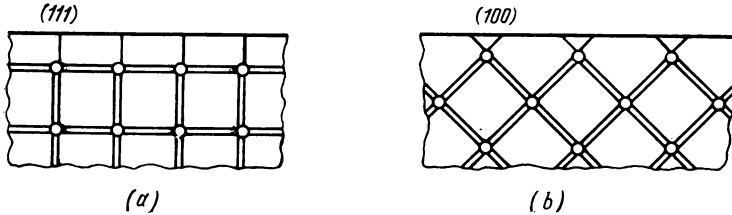


Fig. 2.5. Rupture of covalent bonds on the crystal surface
(a) in plane (111); (b) in plane (100)

directly under the surface, and the dimensions of working regions are often comparable to the thickness of boundary layers. The quantitative features of surface layers will often be noted in the subsequent Sections.

The surface of real crystals commonly has a yet more complex structure than the surface described above. For instance, a film of SiO_2 , which brings about an energy equilibrium on the surface of a silicon slice, may have a hydrate (aqueous) coat that results from adsorption of hydroxyl groups, OH . Such a two-layer film is called a *hydrated oxide* typical for the surface of silicon.

Along with films of physicochemical origin, the surface of a crystal can certainly be **contaminated** with very different substances such as the residues of acids or alkalis utilized for surface treatment, grease spots, and others.

2.2.4. Amorphous substances. Apart from polycrystalline (granular) solids, there are *amorphous*, or completely homogeneous, structureless substances. The distinctive property of amorphous bodies is the absence of a well-defined temperature of melting: transition from the liquid to the solid state proceeds uniformly, so the viscosity grows evenly. Conversion from the solid to the liquid state occurs uniformly too.

A typical representative of amorphous solids is glass of various kinds, including common glass, the glass former of which is silicon oxide, SiO_2 . Most of the thin dielectric films used in microelectronics are amorphous.

To amorphous semiconductors belong chalcogenide glasses, which are the compositions of silicon with chalcogenide elements such as tungsten, tellurium, and others. Amorphous semiconductors are cheaper and easier to produce than single crystals. Besides, they are less subject to radiation-induced defects as are polycrystals. But these materials are still in the stage of development since they show a poor reproducibility and stability of properties and find only limited, specific applications.

2.3. Charge Carriers

One of the most important parameters of any of the substances, semiconductors included, is electric resistivity. It is obvious that a substance can display electric conduction only if it has **free** carriers which can move under the effect of an electric field or gradient of carrier concentration. Consider the origin of free charge carriers in semiconductors.

A pure, perfect semiconductor with an ideal crystal lattice is called *intrinsic*. At absolute zero, such a semiconductor has no free carriers and represents an ideal insulator. As its temperature rises, the crystal acquires an additional energy which induces vibratory motions of atoms located at the lattice sites. In accordance with a corpuscular interpretation, the carriers of energy that induce mechanical vibrations of the lattice are quanta known as *phonons*, which are analogs of photons.

With an increase in the temperature, the number and energy of phonons grow high enough to break off covalent bonds between the atoms of the lattice. The rupture of a covalent bond results in a pair of charge carriers—a free (knock-on) electron and a vacancy, called a *hole*, near the atom which has lost the electron (Fig. 2.6). The process of electron-hole generation under the action of phonons is *thermal generation*.

One of the **valence** electrons of a neighbor atom escapes its bond and quickly fills the hole to complete the broken bond, leaving behind a new hole, and so on. *The hole thus behaves as a particle carrying an elementary positive charge*. Like a free electron, a hole executes a random motion for a certain period of time, called *hole lifetime*, until it recombines with one of the **free** electrons.

It is not only phonons that rupture covalent bonds and generate electron-hole pairs, but also quanta of other kinds of energy, such as light, X-rays, and γ -rays. In comparison with the effect of heat, the specific feature of the last factors is that they only exert a **local** effect which depends on the penetrating power of the energy in question and the area of rays impinging on the surface. In other words, *exposure of a semiconductor to radiation is equivalent to its local heating in the absence of thermal conduction*. If the area of a

beam exceeds the dimensions of a crystal, and the crystal is rather thin (transparent to radiation), the result of irradiation will be in essence the same as in heating.

Semiconductors thus have two types of free carriers, electrons and holes. In an intrinsic semiconductor, they always appear and recombine in pairs, so that the number of electrons is always equal to the number of holes. Conduction in an intrinsic semiconductor, which is due to electron-hole pairs of thermal origin, is termed *intrinsic conduction*. Conduction that results from impurity atoms is

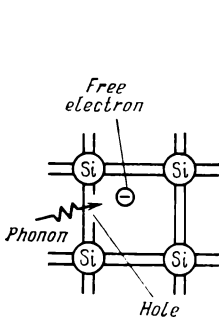


Fig. 2.6. Generation of a hole

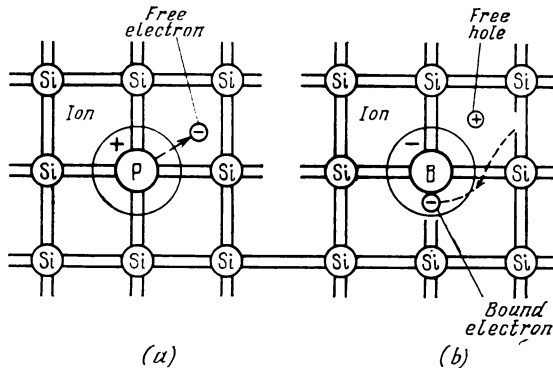


Fig. 2.7. Substitution of impurity atoms for parent crystal lattice atoms

(a) donor impurity (formation of a free electron and a stationary positive ion); (b) acceptor impurity (formation of a free hole and a stationary negative ion)

known as *extrinsic (impurity) conduction*. Impurities inherent in silicon are **substitutional impurities** (see Fig. 2.3c). The result of substitution depends on the valency of impurity atoms.

If we dope silicon with an atom of a **pentavalent** element, say, phosphorus, antimony, or arsenic, four of the five electrons of this atom will interact with four electrons of neighbor host atoms to complete four covalent bonds (Fig. 2.7a) and thus form together a stable shell consisting of eight electrons. The fifth electron of the pentavalent atom is weakly bound to its nucleus and can be readily knocked out by phonons. The impurity atom then becomes an immobile ion with a unit **positive** charge.

Free electrons lost by impurity atoms add to electrons broken away from host atoms. Conduction in the semiconductor becomes largely electronic; such a semiconductor is called *electronic*, or *n type*. Impurities which provide for electronic conduction are known as *donors* since they donate (give off) electrons.

If now silicon receives an atom of a **trivalent** element such as boron, gallium, or aluminium, then all the three valence electrons of

this atom will interact with four electrons of neighbor atoms to build up covalent bonds (Fig. 2.7b). Obviously, one of the four bonds remains incomplete and needs an additional electron to form a stable eight-electron shell. A **valence** electron of a neighbor atom may get free and fill the bond, leaving behind a hole. The impurity atom is now a stationary ion having a unit **negative** charge.

Impurity holes have an added effect on parent holes, and conduction in the semiconductor is largely due to holes. Such a semiconductor is termed a *hole*, or *p-type*¹, semiconductor. Impurities responsible for hole conduction are called *acceptors* since they accept (or capture) electrons from the lattice.

To take off an electron from a donor or an electron from a neighbor host atom to fill the incomplete bond of an acceptor requires a certain amount of energy, called impurity *ionization* or *activation energy*. That is why at absolute zero, ionization does not exist, but at working temperatures ranging from minus 60°C and above, at room temperature in particular, impurity atoms of elements in groups III and V added to silicon and germanium get **ionized almost completely**.

Since in impurity semiconductors the concentrations of electrons and holes differ sharply, it is customary to call the prevailing type majority carriers, and the other type minority carriers. In *n-type* semiconductors majority carriers are electrons, and in *p-type* semiconductors these are holes.

2.4. Energy Levels and Bands

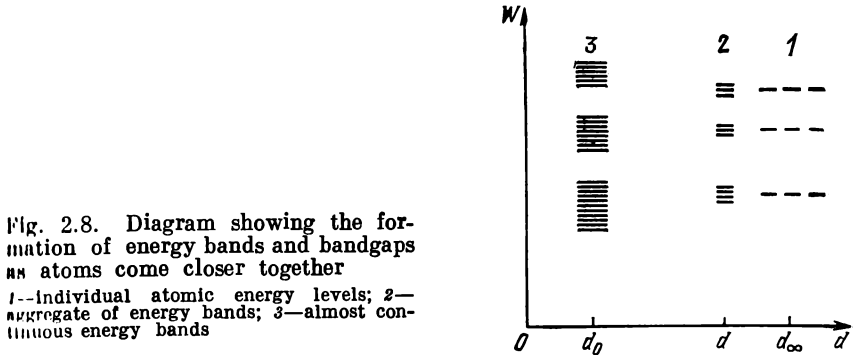
A quantitative analysis of semiconductors and semiconductor devices relies on the band theory of solids.

2.4.1. Band structure of semiconductors. A solid body consists of a host of atoms which *strongly interact* owing to small interatomic distances. Instead of a combination of discrete energy **levels** inherent in an individual atom, characteristic of a solid body is an aggregate of *energy bands*. Every band originates from a certain level which splits, as it were, as atoms come closer together. As a result, a crystal with an interatomic spacing d_0 features a definite *arrangement of energy bands*; the band diagram where (*allowed*) *energy bands alternate with energy gaps*, also called *forbidden bands* or *bandgaps*, appears in Fig. 2.8. The upper energy band is a *conduction band*, and the band below it is a *valence band*. At absolute zero the valence band is always **filled completely** with electrons, whereas the conduction band is

¹ Letters *n* (from negative) and *p* (from positive) are standard symbols accepted in semiconductor physics and technology to identify quantities related to electrons and holes respectively. A symbol *i* (from intrinsic) is used for quantities describing an intrinsic semiconductor.

either filled only in its bottom or empty at all. The first case is specific to metals (Fig. 2.9a), and the second to semiconductors and dielectrics (Fig. 2.9b, c). Dielectrics mainly differ from semiconductors by a much wider bandgap.

Figure 2.9b illustrates the energy band diagram for a pure, intrinsic semiconductor. In extrinsic (impurity) semiconductors, the



band diagrams (Fig. 2.10) are different. As seen, donor and acceptor levels **occupy the bandgap**; donor levels lie close to the lower edge (bottom) of the conduction band (Fig. 2.10a), and acceptor levels

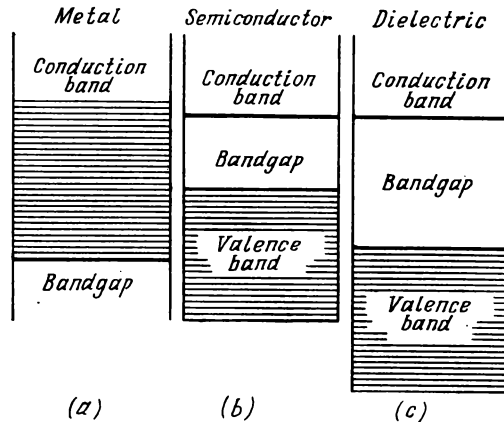


Fig. 2.9. Energy band diagrams at $T = 0$ K
 (a) for metal; (b) for semiconductor; (c) for dielectric

close to the upper edge (top) of the valence band (Fig. 2.10b). It should also be noted that **impurity levels do not split** into bands because the impurity concentration is usually low and, hence, the

distance between impurity atoms is so large that the atomic interaction needed to form energy bands is quite insufficient.

This conclusion does not cover the cases where the impurity density is very high, 10^{18} - 10^{19} cm^{-3} and above. With such impurity concentrations, impurity levels "split" and form an impurity band which commonly combines with the nearest energy band of the semiconductor. The combined band is partially filled with electrons, as is the case with metals. For this reason **highly doped semiconductors** are termed *degenerate*, or *semimetals*.

Donor and acceptor levels are said to be *shallow*, implying that they lie at small distances from respective energy bands. But some impurities present in a semiconductor or specially introduced into it

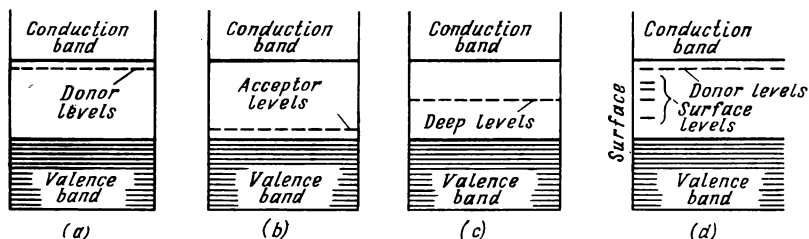


Fig. 2.10. Energy band diagrams for extrinsic semiconductors at $T = 0$ K
(a) *n*-type; (b) *p*-type; (c) neutral impurity doped; (d) *n*-type, including surface states

feature *deep* levels located near the middle of the bandgap (Fig. 2.10c). In silicon, deep levels are typical for the atoms of gold, copper, nickel, and some other impurity elements. Such impurities are neither donors nor acceptors, but they play a significant part in the operation of semiconductor devices. This fact will be given due consideration later in the book.

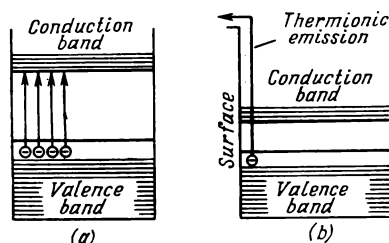
In Subsec. 2.2.3 we have pointed out the influence the surface of a semiconductor has on its various parameters (more precisely, its thin surface layer having a thickness of a few atomic spacings). In the surface layer, crystal lattice defects and adsorbed atoms tend to produce **additional** energy levels, or sometimes even full energy bands. Such specific levels are commonly referred to as *surface levels*, or *surface states*. These levels may occupy any position on the band diagram of a semiconductor. They most often lie within the bandgap as do donor, acceptor, and trapping levels (Fig. 2.10d).

Surface levels are the cause of difference in electrical and physical parameters between the surface layer and crystal bulk. The degree of such a difference depends on the *density of surface states*, N_{ss} . From the physical viewpoint, this parameter defines the quantity of additional levels in the surface layer per unit area.

2.4.2. Carrier transitions between bands and states. In an intrinsic semiconductor, a part of electrons transfer from the valence band to higher levels (at a temperature different from absolute zero), namely, to the conduction band (Fig. 2.11a). The amount of energy required to excite electrons from the valence to the conduction band depends on the width of the bandgap.

Electrons that occupy energy levels in the conduction band are called *free* in the sense that they can move **within** the crystal under the influence of an electric field. For an electron to leave the crystal and escape into the environment, it must gain enough energy to

Fig. 2.11. Transition of valence electrons into the conduction band (a) and out into the environment (b)



overcome a rather high potential barrier on the surface of a solid (Fig. 2.11b). The energy required for an electron to jump over this barrier is known as the *work function*, which determines the *thermionic emission* of a solid at a given temperature. At common operating temperatures, only a very small number of electrons can store this amount of energy. So, thermionic emission has to be reckoned with only in particular cases.

It is possible to liken an aggregate of free electrons in a solid to an electron gas contained, as it were, in a "vessel" formed by the external faces of a crystal. Since such a "vessel" is full of stationary atoms at lattice sites, *the properties of electrons in a solid differ from the properties they have when placed in free space* (vacuum). Thus the mass of an electron in a crystal differs from its mass in a vacuum. That is why the theory of solids commonly utilizes the notion of *effective mass* m^* , which is a few times and occasionally tens of times smaller than the mass of the electron in a vacuum (Table 2.1).

Electrons excited by phonons into the conduction band leave in the valence band empty energy levels, or *holes*. These levels can be filled by electrons in the valence band, the process being equivalent to the **motion** of holes. In Sec. 2.3 we have given a corpuscular interpretation of this process.

The return of an electron from the conduction band onto vacant levels in the valence band results in *recombination* of this electron with a hole, that is, elimination of the pair of charge carriers. In the equilibrium state, the rates of thermal generation and recombination of electron-hole pairs are equal.

Table 2.1

Basic Parameters of Some Semiconductors

Parameter	Silicon	Germanium	GaAs alloy	InSb alloy
Nuclear charge	14	32	—	—
Atomic mass	28.1	72.6	—	—
Relative permittivity	12	16	11	16
Melting point, °C	1 420	940	1 280	520
Thermal conductivity λ , W/cm °C	1.2	0.55	—	—
Specific heat c , J/g °C	0.75	0.41	—	—
Effective electron mass m_n , in relative units	0.33	0.22	0.07	0.013
Effective hole mass m_p , in relative units	0.55	0.39	0.50	0.60
At $T=300$ K				
Bandgap width φ_g , V	1.11	0.67	1.40	0.18
Effective density of states N_c , cm ⁻³	2.8×10^{19}	1.0×10^{19}	—	—
Effective density of states N_v , cm ⁻³	1.0×10^{19}	0.61×10^{19}	—	—
Electron mobility μ_n , cm ² /(Vs)	1 400	3 800	11 000	To 65 000
Hole mobility μ_p , cm ² /Vs	500	1 800	450	700
Intrinsic resistivity ρ_i , Ω cm	To 2×10^5	To 60	To 4×10^8	—
Intrinsic concentration n_i , cm ⁻³	To 2×10^{10}	2.5×10^{13}	To 1.5×10^6	—
Diffusion constant D_n , cm ² /s	36	100	290	To 1 750
Diffusion constant D_p , cm ² /s	13	45	To 12	17
Critical field strength $E_{cr n}$, V/cm	2 500	900	—	—
Critical field strength $E_{cr p}$, V/cm	7 500	1 400	—	—
Maximum velocity $v_{\max n}$, cm/s	10×10^6	6.5×10^6	—	—
Maximum velocity $v_{\max p}$, cm/s	8.0×10^6	6.0×10^6	—	—

In extrinsic semiconductors the process of formation of free carriers at increased temperatures occurs in a different manner. In n -type semiconductors along with the thermal generation of electron-hole pairs, an added process takes place, in which electrons jump from donor levels into the nearest band, the conduction band. In p -type semiconductors, electrons leave the valence band to occupy the nearest acceptor levels. Consequently, excess electrons appear in n -type semiconductors, and excess holes in p -type semiconductors.

As noted earlier, typical donor and acceptor levels are shallow: the activation energy for these levels is by far smaller than that for electrons in the valence band. With a rise in temperature, therefore, the concentration of free carriers that result from the ionization of

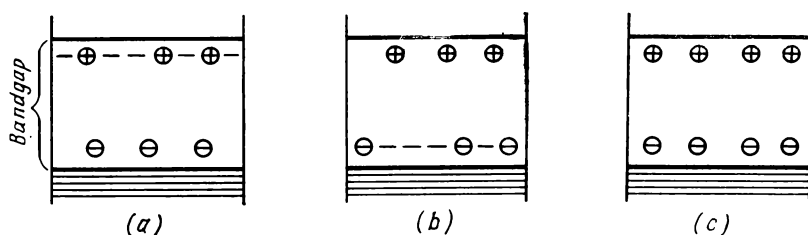


Fig. 2.12. Energy band diagrams for semiconductors containing both types of impurity, at $T = 0$ K

(a) donor impurity prevails; (b) acceptor impurity prevails; (c) concentrations of both impurities are equal

impurity atoms grows much faster than the concentration of electron-hole pairs. The prevailing process of extrinsic carrier generation does not terminate until the ionization of impurity atoms becomes complete. The temperature at which this process ceases is known as the *temperature of complete ionization*. As the temperature rises further, the concentration of free extrinsic carriers remains constant, while the concentration of electron-hole pairs goes on climbing up. This means that the densities of electrons and holes gradually level off with growing temperature, and the *extrinsic semiconductor smoothly changes to an intrinsic semiconductor*. The temperature at which such conversion occurs is called a *critical temperature*.

Real semiconductors generally contain both donor and acceptor impurities, but in different concentrations (N_d and N_a). Band diagrams for such semiconductors are shown in Fig. 2.12. At $N_d > N_a$ (Fig. 2.12a), the number of "useful" donor atoms whose electrons are able to pass to the conduction band is only equal to $N_d - N_a$. The remaining donor atoms give off their "excess" electrons to lower-lying acceptor levels at a temperature near absolute zero, so that the number of negative acceptor ions becomes equal to the number of

positive donor ions. At $N_a > N_d$ (Fig. 2.12b), the number of acceptor atoms ready to capture electrons from the valence band and leave behind holes, is equal to $N_a - N_d$. The rest of acceptor atoms take in electrons from donor atoms at T of about 0 K, with the result that the number of positive donor ions turns out to be equal to the number of negative acceptor atoms.

The differences $N_d - N_a$ and $N_a - N_d$ are called *effective concentrations* of respective impurities. In the further discussion, we shall denote these quantities just as N_d and N_a to avoid the introduction of special designations. Where **total** concentrations are essential, we shall make due reservations.

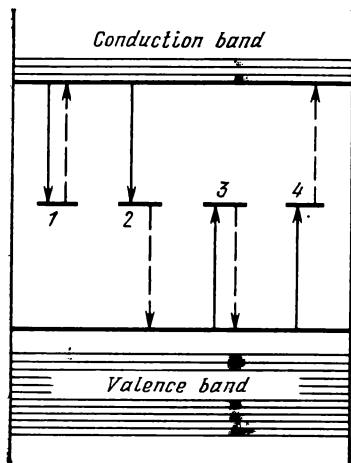


Fig. 2.13. Capture and escape of electrons in a semiconductor having deep impurity levels

If the same amounts of donor and acceptor impurities are introduced into an intrinsic semiconductor, all donor electrons will move to vacant acceptor levels at T slightly above 0 K (Fig. 2.12c). So at any temperature *the concentrations of free carriers in the conduction band and in the valence band of this semiconductor will be the same as in an intrinsic semiconductor*. In the given case, however, in contrast to the intrinsic semiconductor, this semiconductor has a great amount of donor and acceptor ions which have a certain influence on its properties. Semiconductors in which the concentrations of donor and acceptor impurities are equal have received

the name of *compensated semiconductors*.

In conclusion, consider an example of an impurity noted for *deep* levels which lie close to the center of the bandgap (Fig. 2.13). The energy of activation in this case is rather high, so the atoms of such an impurity practically remain unionized and, hence, the concentrations of free carriers do not change. Nevertheless, the role of deep levels may be very substantial. These levels represent the so-called traps, or trap centers, for mobile carriers.

An electron that leaves an energy band and falls into a trap (solid arrows) stays on the trapping level for a definite time, called the *relaxation time*. The electron then either returns to the same band (dash arrows, variants 1 and 3) or moves to another energy band (dash lines, variants 2 and 4). The first process results in a slight temporary change of the number of free carriers, namely electrons in variant 1 or holes in variant 3. The second process involves either

a two-step recombination (variant 2) or a two-step generation of an electron-hole pair (variant 4). There is a much higher probability of two-step than one-step transitions discussed earlier. That is why where the traps are present, the processes of generation-recombination proceed much more intensely and the lifetime of carriers proves much shorter.

The capture of electrons in traps is typical for the surface of a semiconductor which is inherently rich in surface states (see Fig. 2.10d). Depending on the time of relaxation, surface states can be either *fast* or *slow*. For the former states, the relaxation time is in the order of 10^{-8} s, and for the latter, the relaxation time can be around 10^{-3} s and even higher, up to a few seconds.

2.4.3. Characteristic energies and levels. Thermodynamics makes use of the quantity kT to estimate the energy of elementary processes. Here T is the absolute temperature and k is the Boltzmann constant whose value is given in Table 2.2. The quantity kT is close in value to the mean kinetic energy of free electrons ($3/2kT$), which is due to the random motion of these electrons in a solid.

In electronics, the energy of electrons is estimated by the quantity $q\varphi$, where φ is the **potential difference** that an electron passes through, and q is an elementary (electronic) charge whose value is given in Table 2.2. The energy $q\varphi$ is commonly measured in electron-volts

Table 2.2

Basic Physical Constants Used in the Theory of Semiconductors

Quantity	Symbol, value, unit
Elementary charge	$q = 1.602 \times 10^{-19}$ C
Mass of free electron	$m = 9.109 \times 10^{-31}$ kg
Planck constant	$h = 6.62 \times 10^{-34}$ J s
Boltzmann constant	$k = 1.380 \times 10^{-23}$ J K ⁻¹
Electric constant (permittivity of free space)	$\epsilon_0 = 8.854 \times 10^{-12}$ F m ⁻¹
Magnetic constant (permeability of free space)	$\mu_0 = 4\pi \times 10^{-7}$ H m ⁻¹
Avogadro number	$N_A = 6.022 \times 10^{23}$ mol ⁻¹

(eV). An electron-volt is the energy gained by an electron as it passes through a potential difference of 1 volt, and is equal to 1.6×10^{-19} joule. The energy expressed in eV is numerically equal to the corresponding potential difference.

In the theory of semiconductor devices, thermodynamic and electrical processes are closely interlinked. Since it is electrical processes that are of basic practical interest here, we shall express the energy kT in eV, equating it to $q\varphi$. One of the fundamental quantities dealt with in semiconductor physics and engineering follows from the equality $kT = q\varphi$, and is known as *thermal potential*:

$$\varphi_T = kT/q \approx T/11\,600 \quad (2.1)$$

It is useful to remember that at room temperature ($T = 293\text{ K}$), the thermal potential φ_T is approximately equal to 0.025 V or 25 mV.

As with the energy kT , we can express any other energy W on band diagrams in terms of the **energy potential** φ , dividing W by q .

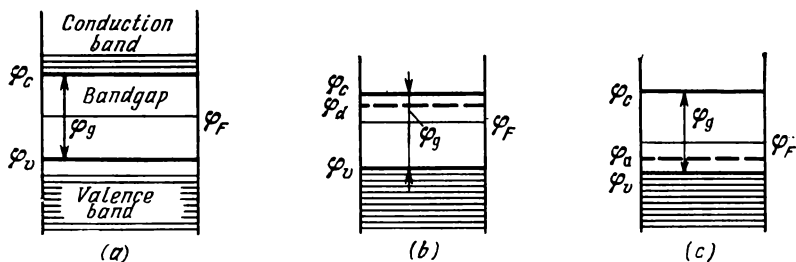


Fig. 2.14. Energy band diagrams for semiconductors at $T \neq 0\text{ K}$
(a) intrinsic; (b) *n*-type; (c) *p*-type

Let us denote the energy levels at the bottom of the conduction band and at the top of the valence band as φ_c and φ_v (where the subscripts *c* and *v* identify respectively the conduction and the valence band), and the difference between these levels—the width of the energy gap—as φ_g (Fig. 2.14):

$$\varphi_g = \varphi_c - \varphi_v \quad (2.2)$$

The *bandgap width* is one of the main parameters of semiconductors: it determines the **energy** required for the generation of electron-hole pairs. The typical values of this parameter lie between 0.2 and 1.5 V (see Table 2.1). The values of φ_g above 2 or 3 V are specific to dielectrics, and below 0.1–0.05 V to semimetals. The bandgap width depends on temperature:

$$\varphi_g = \varphi_{g0} - \varepsilon_g T \quad (2.3)$$

where φ_{g0} is the bandgap width at absolute zero, ε_g (V/°C) is the temperature sensitivity, and T is the absolute temperature. For silicon, $\varphi_{g0} = 1.21\text{ V}$, $\varepsilon_g = 3 \times 10^{-4}\text{ V/°C}$, and, hence, the room temperature value of φ_g is near 1.12 V.

The energy corresponding to the middle of the bandgap is described by the *electrostatic potential* of a semiconductor:

$$\varphi_E = 1/2 (\varphi_c + \varphi_v) \quad (2.4)$$

This potential is often taken as the reference point for other energy potentials.

It should be kept in mind that energy potentials define the energy of **negatively** charged particles (electrons), whereas "classical" electric potentials characterize the energy of **positive** charges. Therefore, any increments and, hence, gradients of electric and energy potentials are opposite in sign; correspondingly, the curves $\varphi_E(x)$ and $\varphi(x)$ are mirror images of each other. This condition should be given due consideration whenever the need arises to estimate the direction of the electric field in a semiconductor or use the Poisson equation.

2.4.4. Distribution of carriers in energy bands. Energy bands contain a huge quantity of levels, from 10^{22} to 10^{23} in 1 cm^3 , and each level **may** have electrons. But the **actual** number of electrons depends on the concentration of donors and temperature (see Subsec. 2.4.2). In order to determine the actual concentration of carriers in a semiconductor, we must know *the distribution of levels and the probability of occupation* of these levels.

For classical semiconductors¹ the occupation probability for a level φ in the conduction band obeys the *Maxwell-Boltzman distribution function*:

$$F_n(\varphi) = \exp \left(- \frac{\varphi - \varphi_F}{\varphi_T} \right) \quad (2.5a)$$

where φ_F is a potential characterizing the Fermi level, which is an electrochemical potential of the system of particles. This potential will be defined later in the book. Formally, the Fermi level is the energy level the occupation probability of which is equal to one half.

The probability that an energy level in the valence band is empty (that there is a hole on this level) is determined by an analogous function:

$$F_p(\varphi) = \exp \left(- \frac{\varphi_F - \varphi}{\varphi_T} \right) \quad (2.5b)$$

Designate the **density** of levels in the conduction band near the level φ as $P(\varphi)$. Then, $P(\varphi) d\varphi$ will be the **quantity** of levels in the

¹ The classical (*nondegenerate*) semiconductor is considered to be a semiconductor having a small impurity concentration that is insufficient for the formation of impurity bands and degeneration of the semiconductor into a semi-metal.

region $d\varphi$. Multiplying this quantity by the occupation probability $F_n(\varphi)$ gives the concentration of free carriers within the range of levels from φ to $\varphi + d\varphi$. The total concentration of free carriers, n , is found by integration over the entire width of the conduction band. If we assume that $P(\varphi) \sim \sqrt{\varphi}$, then

$$n = N_c \exp \left(-\frac{\varphi_c - \varphi_F}{\varphi_T} \right) \quad (2.6a)$$

Here N_c is the effective density of levels (states) in the conduction band:

$$N_c = 0.5 \times 10^{16} (m_n/m)^{3/2} T^{3/2}$$

where m_n is the effective mass of an electron.

In a similar manner, we can obtain the expression for hole concentration:

$$p = N_v \exp \left(-\frac{\varphi_F - \varphi_v}{\varphi_T} \right) \quad (2.6b)$$

Here N_v is the effective density of levels in the valence band:

$$N_v = 0.5 \times 10^{16} (m_p/m)^{3/2} T^{3/2}$$

where m_p is the effective mass of a hole. For silicon, the ratio $N_c/N_v = 2.8$. For simplicity, N_c is often taken equal to N_v .

2.4.5. Intrinsic concentration. Multiplying the left-hand sides and the right-hand sides of Eqs. (2.6) with regard to Eq. (2.2), it is easy to present the product of electron and hole concentrations in the form

$$np = N_c N_v \exp (-\varphi_g/\varphi_T) \quad (2.7)$$

It is clear that at a constant temperature the product of concentrations is constant too: an increase in the concentration of carriers of one type entails a decrease in the concentration of the other type.

In an intrinsic semiconductor, electron and hole concentrations are exactly equal: both are denoted by n_i and called *intrinsic (carrier) concentrations*, or densities. Substituting $n = n_i$ and $p = n_i$ into Eq. (2.7) and finding the square root, we obtain the following expression for intrinsic concentration:

$$n_i = \sqrt{N_c N_v} \exp \left(-\frac{\varphi_g}{2\varphi_T} \right) \quad (2.8)$$

Note two major features of this expression. First, *intrinsic concentration is strongly dependent on the bandgap width*. For example, at $\varphi_{g1} - \varphi_{g2} = 0.2$ V, the intrinsic concentrations will differ by a factor of e^4 , or about 55. It is exactly for this reason that n_i in silicon is three orders of magnitude smaller than in germanium (see Table 2.1). Second, intrinsic concentration is heavily dependent on temperature since φ_T enters the exponent (the effect of temperature

dependence of N_c and N_v is negligible). Thus if $\varphi_g/2\varphi_T = 20$ and the absolute temperature rises by 5% (15°C with respect to room temperature), the intrinsic concentration increases by a factor of e , or about 3 times. It is easy to see that **the effect of temperature increases as the bandgap widens.**

Relation (2.7) is often written in a more compact form, in terms of intrinsic concentration:

$$np = n_i^2 \quad (2.9)$$

2.4.6. Fermi level. Using Eqs. (2.6) with due regard for Eq. (2.4) and assuming for simplicity that $N_c = N_v$, it is not difficult to present the electron-to-hole density ratio in the form:

$$n/p = \exp \left(-\frac{2(\varphi_E - \varphi_F)}{\varphi_T} \right) \quad (2.10)$$

Substitute $p = n_i^2/n$ from (2.9) into the left side of (2.10) and take logarithms of both sides. We are now in a position to express the Fermi level in terms of the free electron concentration:

$$\varphi_F = \varphi_E + \varphi_T \ln (n/n_i) \quad (2.11a)$$

Substituting $n = n_i^2/p$ from (2.9) into (2.10) gives the Fermi level expressed in terms of the hole concentration:

$$\varphi_F = \varphi_E - \varphi_T \ln (p/n_i) \quad (2.11b)$$

The second terms in the right sides of Eqs. (2.11), which characterize carrier densities, are called *chemical potentials*. The Fermi level is thus the sum of the electrical and the chemical potential, hence, its name *electrochemical potential*.

Expression (2.11) allows us to make the following conclusions:

1. In intrinsic semiconductors, where $n = p = n_i$, the Fermi level lies in the **middle** of the bandgap (see Fig. 2.14a).

2. In electronic semiconductors, where $n > n_i$, the Fermi level lies in the **upper** half of the bandgap; the higher the electron density, the higher the position of the Fermi level (see Fig. 2.14b).

3. In hole semiconductors, where $p > n_i$, the Fermi level lies in the **lower** half of the bandgap; the higher the hole density, the closer its position to the upper edge of the valence band (see Fig. 2.14c).

4. As the temperature rises and an impurity semiconductor evenly changes into an intrinsic semiconductor (see Subsec. 2.4.2), the Fermi level shifts to the middle of the bandgap.

One of the fundamental conditions in the physics of semiconductors runs as follows: *the Fermi level is the same in all parts of an equilibrium system whatever its heterogeneity.* This condition may be

expressed in the form of two equipotent relations:

$$\begin{aligned}\varphi_F &= \text{constant}, \\ \text{grad } (\varphi_F) &= 0\end{aligned}$$

The origin of these relations is explained in the next Subsection.

2.4.7. Boltzmann's equilibrium condition. Let the system under discussion be a *p*-type semiconductor consisting of two regions which differ in the concentration of holes. It is obvious that at the boundary (interface) between the two regions there is a concentration gradient and, hence, a chemical potential gradient, with the result that $\text{grad } (\varphi_F) \neq 0$. Such a system is **nonequilibrium**: the concentration gradient will cause holes to diffuse from the region of a higher concentration into the region of a lower concentration.

If holes were neutral particles, the process of diffusion would culminate in the equalization of hole densities. But since holes carry charges, the process develops in another manner.

In the region that takes in holes a positive charge builds up, whereas in the portion that loses holes a negative charge appears; this is due to immobile acceptor ions deprived of holes that balanced out their charge before. This gives rise to an electric field which inhibits the further diffusion of holes. In the end, the gradient of chemical potential becomes offset by the gradient of electrical potential and the resultant gradient of the Fermi level goes to zero.

Such an equilibrium condition at which there exist both the concentration gradient and electric field but their effect is cancelled out and the directional motion of particles does not take place, is known as *Boltzmann's equilibrium condition*.

Boltzmann's equilibrium is particularly evident in heterogeneous semiconductors exhibiting a uniformly distributed impurity concentration. These semiconductors find widespread uses in transistor engineering and microelectronics. Assume that in a *p*-type semiconductor the acceptor concentration smoothly varies along the *x*-axis. The hole concentration will then vary likewise. This means that at every point of the semiconductor there appears a hole concentration gradient dp/dx .

The concentration gradient can only be stationary in the presence of a counteracting electric field. Since the semiconductor does not experience the action of an **external** field, it must have an **internal** electric field. Consequently, *specific to inhomogeneous semiconductors are internal electric fields*. They result from a slight shift of the entire aggregate of majority carriers relative to the aggregate of impurity ions which gave birth to these carriers. In other words, the electric fields are due to *polarization* of inhomogeneous semiconductors; in this case, the basic part of a semiconductor, excluding thin

surface layers, may remain neutral, that is, may be free from bulk (space) charges (Fig. 2.15).

It is easy to determine the internal field strength from the set of Eqs. (2.11). For a p -type semiconductor, differentiate both sides of Eq. (2.11b) with respect to x and allow for the fact that at equilibrium $d\varphi_F/dx = 0$. Then

$$\frac{d\varphi_E}{dx} = \varphi_T \frac{dp/dx}{p}$$

To pass from the energy potential φ_E to the electrical potential φ , we must change the sign of gradient in the left-hand side of the expression (see p. 39). Since $-d\varphi/dx$ is the field strength, we have

$$E = \varphi_T \frac{dp/dx}{p}$$

If the hole concentration **drops** along the x -axis, then $dp/dx < 0$ and the field E becomes negative (known as the brake field), which **retards** holes and thus contributes to establishing Boltzmann's equilibrium.

2.5. Electric Conductivity

The movement of carriers in a semiconductor under the applied field is termed *drift*. The density of drift current is determined by the known expression

$$j = \sigma E \quad (2.12)$$

where σ is conductivity.

Since semiconductors have two types of mobile carrier, conductivity includes two components, an electron and a hole component:

$$\sigma = qn\mu_n + qp\mu_p \quad (2.13)$$

where μ_n and μ_p are the *mobilities* of respective carriers.

The main component in Eq. (2.13) is that which is associated with **majority** carriers. The minority carrier component is generally insignificant. In an intrinsic semiconductor both components are almost equal in value.

In order to evaluate conductivity and thus determine drift current, it is first necessary to know the **concentrations** of electrons and holes.

2.5.1. Carrier concentration. There would seem to be an easy task to find the values of n and p using Eqs. (2.6). But for this we must

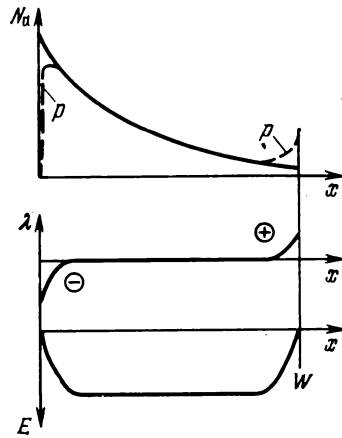


Fig. 2.15. Distribution of concentrations N_a and p , charge density λ , and electric field E in an inhomogeneous p -type semiconductor

know the position of the Fermi level in the bandgap. Meanwhile, the Fermi level is a **function** of carrier concentration as is the chemical potential. So, before calculating the carrier concentration, we must determine the position of the Fermi level. This problem can be solved proceeding from the guiding *condition of neutrality*, which reads: *in a homogeneous (or quasihomogeneous) semiconductor substantial bulk charges are nonexistent whether the semiconductor is in equilibrium or exposed to an externally applied field.*

The condition of neutrality is based on the phenomenon of *dielectric relaxation* which can be illustrated by the following examples. Assume that a local **negative** charge has appeared in an electronic semiconductor. The buildup of such an electronic charge entails the creation of a "clean-out" field which **disperses** the bunch of electrons instantly, in the order of 10^{-12} s. Let us now assume that a local **positive** charge had developed in the same semiconductor. The internal field produced in the semiconductor will cause electrons to move from neighboring regions into the charge region and **neutralize** the positive charge in the same length of time, 10^{-12} s. Consequently, in none of the cases can the bulk charge exist for a more or less long period of time. Based on the condition of neutrality, we may write the following relationship for an *n*-type semiconductor:

$$n = N_d^* + p \quad (2.14)$$

where N_d^* is the concentration of positive donor ions (keeping in mind that this is an **effective** concentration, see Subsec. 2.4.2). Expressing the hole concentration through the electron concentration with the aid of (2.9) and solving the resultant quadratic equation for n , we derive the equation for electron concentration:

$$n_n = \sqrt{\left(\frac{N_d^*}{2}\right)^2 + n_i^2} + \frac{N_d^*}{2} \quad (2.15a)$$

In a similar manner, we can determine the hole concentration in a *p*-type semiconductor:

$$p_p = \sqrt{\left(\frac{N_a^*}{2}\right)^2 + n_i^2} + \frac{N_a^*}{2} \quad (2.15b)$$

Subscripts n and p identify semiconductors with the respective types of conductivity.

The lower operating temperature limit of extrinsic semiconductors is the temperature of complete impurity ionization (from minus 70°C to minus 100°C for silicon) and the upper temperature limit is the critical temperature at which an extrinsic semiconductor becomes intrinsic (see p. 35). For this temperature range, the set of formulas (2.15) can be simplified if we replace the effective concentration of impurity ions, N^* , by the effective concentration of impurity

atoms, N (since over the operating range, practically all impurity atoms are ionized), and neglect the intrinsic concentration n_i (since over the operating range, n_i is substantially lower than the impurity concentration). The concentration of **majority** carriers can then be written in the form

$$n_n = N_d \quad (2.16a)$$

$$p_p = N_a \quad (2.16b)$$

The concentration of minority carriers is easy to determine resorting to relationship (2.9)

$$p_n = n_i^2/N_d \quad (2.17a)$$

$$n_p = n_i^2/N_a \quad (2.17b)$$

The critical temperature T_{cr} can be found from Eq. (2.8) by setting n_i on the left of the expression equal to aN , where N is the impurity concentration. Taking the logs of both sides of the equation and using (2.1), we get

$$T_{cr} = 5800 \frac{\varphi_g}{\ln(\sqrt{N_c N_v}/aN)} \quad (2.18)$$

Since the effective densities of states, N_c and N_v , depend on temperature, as is obvious from Eqs. (2.6), the value of T_{cr} has to be found by the method of successive approximations, starting with the calculation at room temperature (N_c and N_v at room temperature are given in Table 2.4).

As an example, set a at 0.1, N at 10^{16} cm^{-3} , and take the remaining parameters for silicon from Table 2.4. In the second approximation we then find that T_{cr} is equal to about 330°C. The value of T_{cr} (345°C) close to the calculated value can be determined by a semi-empirical formula

$$T_{cr} = 273 \left(\frac{10}{4.5 + \log \rho} - 1 \right) \quad (2.19)$$

where ρ is equal to $0.85\Omega \text{ cm}$ at $N = 10^{16} \text{ cm}^{-3}$.

It is easy to see that T_{cr} depends on impurity concentration: the greater the impurity concentration, the higher the critical temperature. Besides, T_{cr} rises with an increased bandgap width. Indeed, the larger the value of φ_g , the smaller the intrinsic concentration n_i [see Eq. (2.8)]; hence, a higher temperature is necessary for n_i to become equal to extrinsic concentration.

This condition offers an explanation of why *wide-gap semiconductors* attract much interest. A GaAs semiconductor, for example, can in principle make a suitable material for use in the fabrication of integrated elements. Its bandgap width φ_g is 25% larger than that in silicon (see Table 2.4). This means that the *absolute* critical temperature for this material will be 25% higher, all other conditions

being equal. Thus at $N = 10^{16} \text{ cm}^{-3}$ we find that T_{cr} is equal to about 540°C . Despite this important advantage, however, gallium arsenide has not become an alternative as regards the choice between this material and silicon. The fact is that the choice of a material for ICs depends on many characteristics rather than on one parameter only.

From Eq. (2.17) it follows that at low temperatures the concentration of minority carriers is very small. Thus if $N_d = 2 \times 10^{17} \text{ cm}^{-3}$, then at room temperature the hole density in silicon, according to Eq. (2.17a), is merely $2\,000 \text{ cm}^{-3}$, that is, 14 orders of magnitude less (!) than the electron density. But as the temperature increases, the minority-carrier density grows very fast, in proportion to n_i^2 , that is, incomparably faster than even the intrinsic-carrier density. Thus the growth of temperature by 50°C causes an increase in the minority-carrier density by approximately 3 orders of magnitude.

The other factors such as light and various kinds of ionizing radiation have the same strong effect on the concentration of minority carriers. Therefore in semiconductor devices and IC elements whose operation depends on *minority* carriers, these factors should be eliminated wherever possible. On the other hand, the effect of these factors can be turned to proper use in special applications, for example, in photosensitive devices and radiation dosimeters.

2.5.2. Carrier mobility. As known, in free space electrons execute a **uniformly accelerated** motion on exposure to an electric field.

In a solid, moving electrons continuously collide with crystal lattice sites, impurities, and defects, so they undergo what is called *scattering*. A uniformly accelerated motion of electrons in an electric field is only possible in short time intervals between collisions, when the electrons cover a *free path*. After every collision, an electron must, roughly speaking, gather speed anew. The **mean** drift velocity of electrons and holes proves to be quite a definite value, proportional to field strength:

$$\bar{v} = \mu E$$

The proportionality factor μ is the *mobility* of carriers, measured in $\text{cm}^2/\text{V s}$. At a field strength of 1 V/cm , the mobility is numerically equal to the velocity of carriers.

Since there is a difference in effective mass between the electrons and holes, the mobilities of these two types of carriers are different. *The mobility of electrons is as a rule higher than that of holes.* In silicon, electrons have about 3 times the mobility of holes (see Table 2.1). The greater the mobility, or velocity of carriers, the higher the speed of response of a semiconductor device.

This explains why materials that surpass silicon in carrier mobility arouse intense interest. But here too, as in the case of wide-band

semiconductors, one should avoid a biased approach when estimating the usefulness of such materials. For example, the electron mobility in indium antimonide, InSb, is tens of times higher than the electron mobility in silicon (see Table 2.1). Nevertheless this material is unsuitable for ICs since it has a narrow bandgap and thus a rather low critical temperature, which in some cases drops below room temperature.

Carrier mobility is a function of some factors, the most important being temperature, impurity concentration, and field strength. These factors should be given due consideration in the development of semiconductor devices and integrated elements.

The dependence of carrier mobility on temperature is determined by a particular mechanism of carrier scattering. If the prevailing mechanism of scattering is due to lattice vibrations, then

$$\mu_L = \mu_{0L} (T_0/T)^c \quad (2.20a)$$

If the mechanism of scattering due to impurity ions prevails,

$$\mu_I = \mu_{0I} (T/T_0)^{3/2} \quad (2.20b)$$

Here, μ_0 relates to an initial temperature T_0 , for example, room temperature; μ relates to an arbitrary temperature T , the subscripts L and I refer respectively to lattice and impurity mobilities, and the exponent c is a coefficient dependent on the material proper and type of conductivity. For n -type and p -type silicon, c is equal to about 5/2.

The resultant mobility is close to the lower of the two components,

μ_L and μ_I . For silicon, μ_L proves smaller than μ_I at $T > 0^\circ\text{C}$, so the function, $\mu(T)$, is described by formula (2.20a): *carrier mobility decreases with rising temperature*. At $T < \text{minus } 50^\circ\text{C}$, the component μ_I turns out to be smaller than μ_L , and thus the function, $\mu(T)$, is described by formula (2.20b): *carrier mobility drops off with decreasing temperature*. Over the operating temperature range from minus 60°C to plus 125°C the mobility of carriers may vary in magnitude by a factor of 4 or 5, which is of course substantial. The mobility-impurity concentration relation is complex (Fig. 2.16) and on the whole does not lend itself to analytic description. Of the two mobility components μ_L and μ_I , the first is independent of impurity concentration, while the second is proportional to N^{-1} . Therefore, with low impurity concentrations ($N < 10^{15} \text{ cm}^{-3}$), in which case $\mu_I \gg \mu_L$, the resultant

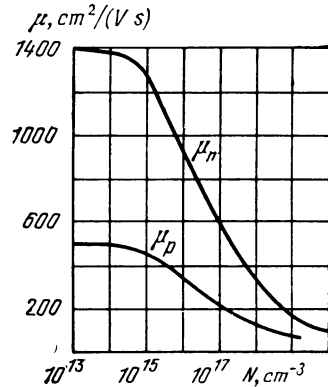


Fig. 2.16. Carrier mobility in silicon versus impurity concentration at $T \approx 20^\circ\text{C}$

mobility is determined by the component μ_L , so that the μ - N relation may be disregarded. In the most important region ($N > 2 \times 10^{15} \text{ cm}^{-3}$), both components remain first comparable in value, then the component μ_T becomes lower and thus determines the resultant mobility. For this range of concentrations, two approximations are valid:

$$\mu = \mu_0 (N_0/N)^{1/3} \quad (2.21a)$$

or

$$\mu = \mu_0 - \Delta\mu \log (N/N_0) \quad (2.21b)$$

In both approximations the values of μ_0 correspond to **tabulated** data (that is, to **low** impurity concentrations) and the value of N_0 is approximately equal to $2 \times 10^{15} \text{ cm}^{-3}$. The coefficient $\Delta\mu$ in formula (2.21b) represents a change in mobility with an increase in concentration by a factor of 10. For n -type and p -type silicon, it is safe to assume that

$$\Delta\mu_n = 400 \text{ cm}^2/\text{V s}, \quad \Delta\mu_p = 200 \text{ cm}^2/\text{V s}$$

As is apparent from Fig. 2.16, the carrier mobility may change about tenfold in the range of impurity concentration from 10^{15} to 10^{19} cm^{-3} . This fact has particular significance in **inhomogeneous** semiconductors extensively used in ICs. However, the inclusion of the function $\mu(N)$ in calculation formulas makes the equations nonlinear and hardly suitable for practical purposes. Therefore, for inhomogeneous semiconductors, one has to **average** the carrier mobility one way or another with the aid of formulas (2.21) and use constant **averaged** values in calculations.

The carrier mobility-field relationship plays a specific part since, according to Eq. (2.12), it affects the current density-field strength relationship which turns nonlinear and, hence, the volt-ampere characteristic of a semiconductor will be nonlinear too. In other words, *the function, $\mu(E)$, leads to a violation of Ohm's law in semiconductors.*

Theory and experiment show that in **weak** fields, where the field strength is lower than a certain critical value ($E < E_{cr}$), the carrier mobility remains constant. The values of E_{cr} are given in Table 2.1. In **supercritical conditions**, at which $E > E_{cr}$, the mobility depends on field strength in the following manner:

$$\mu = \mu_0 (E_{cr}/E)^{1/2} \quad (2.22)$$

Here μ_0 corresponds to the critical field strength, that is, this value of mobility is a *nominal* value.

From the physical standpoint, *the critical field strength corresponds to a condition at which the transport (drift) velocity of carriers becomes*

comparable to their random (thermal) velocity¹. That is why the “total” carrier velocity under supercritical conditions proves higher than the thermal velocity and, hence, the temperature of carriers is higher than that of the semiconductor and environment.

Carriers having an increased temperature, that is, an energy comparable to or exceeding a thermal energy of $3/2kT$ are called *hot*². Since the energy kT is characterized by the thermal potential [see Eq. (2.1)] which reaches 25 mV at $T \approx 300$ K, carriers with an energy of 0.03 to 0.04 eV and above can be regarded as hot carriers at room temperature.

On colliding with phonons, hot carriers raise phonon energy and thus “heat up” the crystal lattice. This results in a new phenomenon called *carrier velocity saturation* or in current saturation at sufficiently high voltages. In this case the speed of carriers no longer depends on field strength, and so the condition $\mu(E) = \text{constant}$ becomes viable. Correspondingly, the μ - E relation takes the form $\mu \sim E^{-1}$. Carrier velocity saturation generally occurs at $E > 4E_{cr}$.

The maximum, or limiting, speed of carriers, v_{\max} , is close to the mean thermal speed v_T and averages 10^7 cm s⁻¹. More accurate values of v_{\max} appear in Table 2.1.

2.5.3. Conductivity. Rewrite common expression (2.13) for intrinsic and extrinsic semiconductors.

In an intrinsic semiconductor, $n = p = n_i$, and, hence,

$$\sigma_i = qn_i (\mu_n + \mu_p) \quad (2.23)$$

Taking from Table 2.1 the values of intrinsic-carrier concentration and mobility for silicon at room temperature, we get $\sigma_i \approx 6 \times 10^{-6} \Omega^{-1} \text{ cm}^{-1}$ and $\rho_i \approx 200 \text{ k}\Omega \text{ cm}$. At this resistivity, a 1 by 1 mm² intrinsic silicon wafer 0.3 mm thick will have a “transverse resistance” (**normal** to the surface) of about 0.5 M Ω and “lateral resistance” (**along** the surface) of some 5 M Ω .

The intrinsic conductivity-temperature dependence is determined by the relation between the intrinsic concentration n_i and temperature [see Eq. (2.8)]. As noted earlier, this dependence is very strong and takes on an exponential form. Fig. 2.17a illustrates the function, $\sigma_i(T^{-1})$ for silicon plotted to a semilog scale; it also shows the scale in degrees Celsius. It is seen that over the operating range from minus 60°C to plus 125°C the intrinsic conductivity of silicon varies

¹ The mean thermal speed of particles in a solid is expressed as $v_T = \sqrt{3kT/m^*}$, where m^* is the effective mass. If we assume that $m^* = m$ and $T = 300$ K, then $v_T \approx 10^7$ cm s⁻¹.

² A small quantity of hot electrons are always present in a solid body even in the absence of an electric field in view of the statistical distribution of electrons according to their energies. In supercritical conditions, the electrons with a mean energy are hot.

by 5 orders of magnitude. In materials having a narrower energy gap (in germanium, for example), variations in σ_i are lower, though the values of σ_i by themselves will be greater because of a rather large intrinsic concentration [see description of Eq. (2.8)].

Disregarding the terms related to minority carriers in (2.13) and using the set of Eqs. (2.16), for n -type and p -type extrinsic semiconductors, we get

$$\sigma_n = qN_d\mu_n \quad (2.24a)$$

$$\sigma_p = N_a\mu_p \quad (2.24b)$$

Over the working temperature range, concentrations N_d and N_a may be assumed constant. In this temperature range, then, the

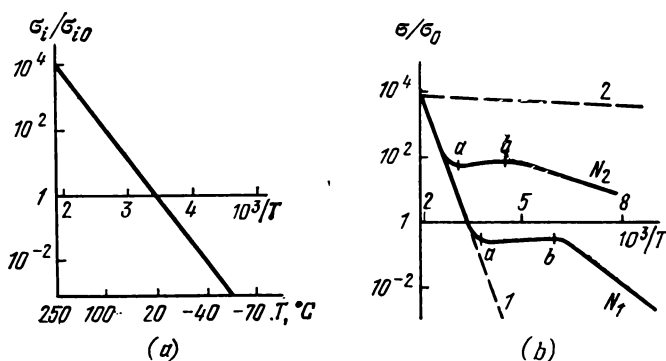


Fig. 2.17. Relative conductivity of silicon *versus* temperature; σ_{i0} and σ_0 are conductivities at $+20^\circ\text{C}$

(a) intrinsic silicon; (b) extrinsic silicon

conductivity-temperature relation for an **extrinsic** semiconductor will be determined by the carrier mobility-temperature relation [see Eqs. (2.20)].

Figure 2.17b displays two plots of σ/σ_0 against T^{-1} at various impurity concentration ($N_2 > N_1$) and also shows for comparison a dashed curve *1* which is a part of the function $\sigma_i (T^{-1})$ presented in Fig. 2.17a. Points *a* represent **critical** temperatures [see Eq. (2.18)], at which the extrinsic semiconductor becomes intrinsic, and therefore to the left of points *a* the curves σ pass into the dashed curve σ_i . Points *b* correspond to the temperature of impurity ionization: to the right of these points (that is, at lower temperatures), the concentration of ionized impurity atoms diminishes and thus conductivity decreases.

As seen, in the working range the conductivity of extrinsic semiconductors is much more weakly dependent on temperature than

that of intrinsic semiconductors. Besides, the temperature dependence of extrinsic conductivity shows a "reverse" character: as the temperature rises, σ decreases rather than grows.

The dashed curve 2 is a plot of σ versus T^{-1} for highly doped, **degenerate** semiconductors, where σ varies very weakly with temperature. This attests to affinity between these semiconductors and metals and explains why they received the name **semimetals**.

2.6. Carrier Recombination

The processes of generation and recombination are inseparable from each other, though opposite in character. Recombination counteracts the process of carrier accumulation and establishes equilibrium carrier densities [see Eqs. (2.6)]. Also, recombination underlies fundamental relations (2.7) and (2.9) and determines the finite *lifetime* of carriers—the parameter which is largely responsible for the transient time.

2.6.1. Recombination mechanisms. The main types of recombination are *direct recombination* and *indirect recombination via impurity centers*.

Direct recombination, or band-to-band recombination, is the transition of an electron across the bandgap from the conduction band to the valence band, where it occupies one of the empty levels and thus "eliminates" a hole. As it makes such a transition, the electron must certainly give up an energy $q\varphi_g$ expended previously on its transfer from the valence to the conduction band. The electron releases this energy in the form of either a photon (the process being known as *radiative recombination*) or a phonon (*nonradiative recombination*). In most semiconductors, silicon included, the probability of radiative recombination is a few orders of magnitude smaller than that for nonradiative recombination. The reason is that as an electron drops from the conduction band into the valence band, it must give up not only its energy but also its **momentum**. Since a photon is unable to accept any amount of momentum whatever, it is necessary that a third particle—phonon—should come into action, but such a combined process takes place on extremely rare occasions.

However, the probability of nonradiative **direct** recombination in itself is very small too, since the comparatively large energy $q\varphi_g$ (near 1 eV) can rarely convert into a single phonon, while its simultaneous distribution between two phonons is hardly probable. Thus *direct recombination as a whole cannot be the main mechanism of recombination in semiconductors*.

It is recombination via impurity centers that plays the main part in the recombination process. Under impurity centers we mean deep levels, or *traps*, located near the center of the bandgap (see Fig. 2.13).

This is a consecutive type of recombination that occurs in **two steps**. First, an electron falls from the conduction band into a trap level, and then down to the valence band. At each step the electron gives off an energy approaching $1/2q\phi_g$ rather than the entire energy as is the case in direct recombination. This sharply raises the probability of energy transfer to a phonon and thus explains why this mechanism plays a prevailing role.

Along with impurity atoms, various lattice **defects** can act as traps. For this reason, an increased rate of recombination is particularly specific to polycrystals, in which all faces between individual grains are defects, and to surface layers of any single crystal semiconductor, where violations of the periodicity of a lattice and broken covalent bonds are inevitable.

2.6.2. Equilibrium carrier recombination. The probability that an electron will directly recombine with **one** of the holes may be written in the form

$$r = \sigma_{ef} v_T$$

where σ_{ef} is the *effective capture cross section*, and v_T is the mean thermal velocity of electrons¹. The quantity r is called a *recombination coefficient* (proportionality constant). Multiplying the coefficient r by the hole density, we find the total probability of recombination of an electron in a unit time with **any** of the available holes. The reciprocal quantity is a mean time interval between the acts of recombination, that is, the *mean lifetime* of electrons in the direct recombination process:

$$\tau_n = 1/rp_0 \quad (2.25a)$$

By reasoning in an analogous way, we determine the mean lifetime for holes:

$$\tau_p = 1/rn_0 \quad (2.25b)$$

In formulas (2.25) and elsewhere in this book the subscript "0" identifies equilibrium conditions.

If we multiply the probability of recombination of **one** electron, rp_0 , by the electron concentration n_0 , we shall obtain the total number of acts of recombination per unit time, that is, *the rate of direct recombination*:

$$R_0 = rn_0p_0 \quad (2.26)$$

¹ A stationary electron will obviously never run into a hole; the higher the speed of an electron, the higher the probability of its "meeting" a hole. As regards the capture cross section, this is a criterion which characterizes the volume around a hole from which an electron, once it has fallen into this volume, is unable to escape; the hole will thus inevitably draw the electron despite its inertia of motion.

From Eqs. (2.25) it is seen that the equilibrium lifetimes for electrons and holes **sharply differ** in the general case because of the difference between concentrations n_0 and p_0 . Besides, the lifetime of **minority** carriers is always **shorter** than that of majority carriers.

Replacing rn_0 in the right-hand side of Eq. (2.26) by $1/\tau_p$ or rp_0 by $1/\tau_n$, we can write the recombination rate in one more widespread form:

$$R_0 = p_0/\tau_p = n_0/\tau_n \quad (2.27)$$

2.6.3. Direct recombination. In a semiconductor found to be in a nonequilibrium state, free carrier concentrations differ from equilibrium-carrier values:

$$n = n_0 + \Delta n \quad (2.28a)$$

$$p = p_0 + \Delta p \quad (2.28b)$$

Nonequilibrium-carrier concentrations n and p may be larger or smaller than equilibrium concentrations (and thus signs affixed to Δn and Δp in Eqs. (2.28) may be both positive and negative). Increments Δn and Δp are called *excess concentrations*, or densities. For a semiconductor to be electrically neutral, excess concentrations of electrons and holes must be equal:

$$\Delta n = \Delta p \quad (2.29a)$$

Moreover, with variations in excess concentrations, electrical neutrality must also remain unchanged. Hence, the equality condition for the rate of changes in concentration:

$$dn/dt = dp/dt \quad (2.29b)$$

From Eqs. (2.29) it follows that *there is no sense in analyzing separately the behavior of excess electrons and excess holes* because the functions $\Delta n(t)$ and $\Delta p(t)$ are identical. In the further discussion, therefore, we shall consider the behavior of excess electrons only.

Suppose that for some reason or other the equality between the rates of generation and recombination has become upset. The electrons then will be stored up (or swept out) with a rate equal to the difference between the rates of generation and recombination:

$$dn/dt = g - r(np) \quad (2.30)$$

where g is the generation rate, and $r(np)$ is the recombination rate.

Write the generation rate in the form

$$g = g_0 + \Delta g = r(n_0 p_0) + \Delta g \quad (2.31)$$

where g_0 is the equilibrium value equal to the equilibrium recombination rate [see Eq. (2.26)].

Transform the recombination term in the right side of (2.30) by introducing Eqs. (2.28) and taking account of Eqs. (2.29):

$$r(np) = r[n_0 p_0 + \Delta n(n_0 + p_0) + \Delta n^2]$$

By assuming $\Delta n \ll n_0 + p_0$, we have ground to omit the term Δn^2 and thus linearize Eq. (2.30). Further, express the concentrations n_0 and p_0 in parentheses through lifetimes, using Eqs. (2.25). We get

$$r(np) = r(n_0 p_0) + \Delta n \left(\frac{1}{\tau_n} + \frac{1}{\tau_p} \right)$$

Last, introduce the *equivalent lifetime* τ in the form of a relation

$$\frac{1}{\tau} = \frac{1}{\tau_n} + \frac{1}{\tau_p} \quad (2.32)$$

The recombination rate will now take the form

$$r(np) = r(n_0 p_0) + \Delta n / \tau$$

Substituting into Eq. (2.30) the resultant expression of $r(np)$ and also that of the generation rate g from Eq. (2.31), we obtain the *accumulation equation* of excess carriers in the form

$$dn/dt = \Delta g - \Delta n / \tau \quad (2.33)$$

Setting $\Delta g = 0$, we get the *recombination equation*

$$dn/dt = -\Delta n / \tau \quad (2.34)$$

Solving the recombination equation gives the exponential function

$$\Delta n(t) = \Delta n(0) \exp(-t/\tau) \quad (2.35)$$

where $\Delta n(0)$ is the initial excess concentration. Relation (2.35) permits us to determine the lifetime as *a time interval during which the excess concentration decreases by a factor of e*.

From the structure of Eq. (2.32) it follows that the quantity τ approaches the **minimum** of its two components τ_n and τ_p . So, *the equivalent lifetime of excess carriers is determined by the lifetime of minority carriers*. In *n*-type semiconductors, $\tau \approx \tau_p$, and in *p*-type semiconductors, $\tau \approx \tau_n$.

2.6.4. Trap recombination. With the **trap mechanism** of recombination, the recombination rate of excess carriers is described by the *Shockley-Read formula*:

$$\frac{dn}{dt} = - \frac{np - n_0 p_0}{(n + n_t) \tau_p + (p + p_t) \tau_n} \quad (2.36)$$

Here n_t and p_t are parameters having the dimensions of concentration, dependent on the distribution of trap levels in the bandgap¹;

¹ The concentrations n_t and p_t are a few orders of magnitude lower than the concentration of majority carriers, but may be tens of times higher than the **intrinsic** concentration.

and τ_p and τ_n are minority carrier lifetimes:

$$\tau_n = 1/r_n N_t \quad (2.37a)$$

$$\tau_p = 1/r_p N_t \quad (2.37b)$$

where N_t is the trap center concentration.

Formula (2.36) has the same structure as Eqs. (2.25). In direct recombination, however, the lifetimes were different due to the difference in carrier concentration, but in the given case they differ on account of the difference between the recombination coefficients.

Equating the right sides of Eq. (2.36) and Eq. (2.34), it is easy to determine the lifetime τ . Substituting the values of n and p from Eqs. (2.28) with regard to Eq. (2.29a) and assuming $\Delta n \ll n_0 + p_0$ [as has been done in deriving (Eq. 2.33)], we get

$$\tau = \frac{n_0 + n_t}{n_0 + p_0} \tau_p + \frac{p_0 + p_t}{n_0 + p_0} \tau_n \quad (2.38)$$

For an electronic semiconductor, $\tau \approx \tau_p$, as follows from Eq. (2.38), if inequalities $n_0 \gg p_0$ and $n_0 \gg n_t$, p_t characteristic of this semiconductor type, hold good. For a p -type semiconductor, the conditions being the same, $\tau = \tau_n$. This means that as with direct recombination, in the case of trap recombination, *the lifetime of excess carriers depends on the lifetime of minority carriers.*

The lifetime-trap density relation follows from expressions (2.37): the higher the density of trap centers, the smaller the lifetime.

To illustrate the relation between the lifetime and impurity concentration, we consider an example of an n -type semiconductor in which the lifetime is determined by the first term on the right of Eq. (2.38):

$$\tau \approx \frac{n_0 + n_t}{n_0 + p_0} \tau_p \quad (2.39)$$

If the donor concentration is sufficiently high, $n_0 \gg p_0$, n_t . The lifetime is then independent of impurity concentration: $\tau \approx \tau_p$. As the donor density falls off, the inequality $n_0 \gg n_t$ becomes upset and the lifetime grows. In the limit, when the donor density approaches zero, the semiconductor goes intrinsic and the lifetime reaches a maximum:

$$\tau \approx \frac{n_t}{2n_i} \tau_p \gg \tau_p \quad (2.40)$$

Similar results can be obtained for a p -type semiconductor. The behavior of the function $\tau(N)$ is illustrated in Fig. 2.18a. As seen, *in strongly doped semiconductors the carrier lifetime is shorter than in weakly doped and intrinsic semiconductors.*

The lifetime variation with temperature is attributed to a sharp growth of the concentration n_t by the law congruent to Eq. (2.8).

As the quantity n_t becomes comparable to n_0 , the lifetime starts to grow, and at $n_t > n_0$ the function $\tau(T)$ practically coincides with the exponential function $n_t(T)$. The growth in lifetime is no longer so fast as before near the critical temperature when the semiconductor turns to intrinsic. For an intrinsic semiconductor, the function $\tau(T)$, as is clear from Eq. (2.40), has a decreasing rather than an increasing

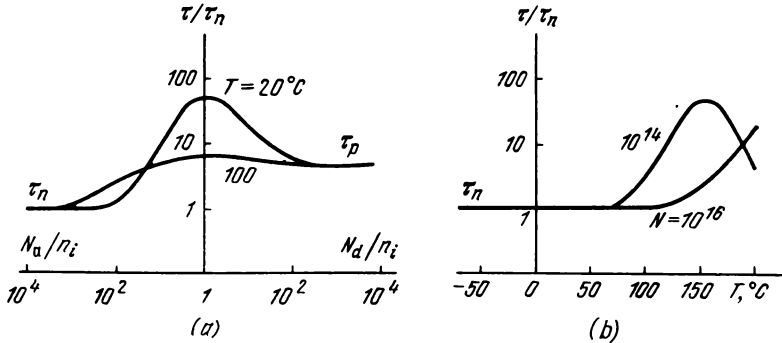


Fig. 2.18. Carrier lifetime versus impurity concentration (a) and temperature (b)

character, because the concentration n_t rises slower than n_i . Fig. 2.18b displays examples of the function $\tau(T)$ at various concentration levels. It is seen that the temperature dependence of lifetime is most noticeable for slightly doped semiconductors (for silicon with $N < 10^{14}\text{--}10^{15}\text{ cm}^{-3}$). In heavily doped (low-resistance) semiconductors, this dependence is of the secondary importance.

It should be noted in conclusion that the typical lifetimes for silicon lie between 0.1 and 1 μs . In silicon specially doped with a trap impurity, commonly gold, the lifetime falls to 10 ns and below.

2.6.5. Surface recombination. The recombination processes in the surface layer of a semiconductor do not in principle differ from those taking place in its bulk. But the surface is noted for a **specific band structure** (see Fig. 2.10d) and, hence, has **quantitatively** other parameters than the bulk. This fact should not be ignored in analyzing and designing semiconductor devices and ICs, all the more so, as the active regions of ICs lie near the surface (see Fig. 1.3). Let us denote the surface lifetime as τ_s , and the bulk (volume) lifetime as τ_v .

If the working portion of an integrated element is entirely located in the surface layer or in the crystal bulk, then in the analysis we must use the parameter τ_s or τ_v respectively. If however the working region "comes out to the surface", as is often the case, and thus lies partly in the surface layer and partly in the bulk, it is customary to

utilize the *effective lifetime* τ , the expression for which has the form

$$1/\tau = 1/\tau_s + 1/\tau_v \quad (2.41)$$

It is this parameter that serves as a lifetime criterion in the analysis of transistors and other integrated elements.

Since τ_s is generally smaller than τ_v due to a high trap density near the surface, the effective lifetime more approximates τ_s . The latter quantity, however, is more difficult to calculate and measure than τ_v . For this reason another parameter, called the *surface recombination velocity* s (cm s⁻¹), has found widespread use for evaluating surface recombination, because this quantity is easier to measure than the lifetime τ_s . The surface recombination velocity is heavily dependent on the method and quality of crystal surface treatment. Its typical values lie in the range from 100 to 10⁴ cm s⁻¹ and above.

The physical meaning of the parameter s is the following. If an excess concentration of carriers appears near the surface layer, where the recombination rate is higher than in the bulk, then a major portion of excess carriers will move to the surface to make up for the loss of carriers in the surface layer. Consequently, a *flow of carriers appears* between the bulk and surface, the speed of this flow being just the parameter s .

In the general case, it is difficult to establish the relation between the surface recombination velocity s and the surface lifetime τ_s . It has been possible to solve the problem only for two individual, though important cases: for a bar of infinite length and for a thin plate of thickness d and infinite area. For the latter case, most interesting from the practical point of view, the following relation holds:

$$\tau_s = d^2/4\eta D \quad (2.42)$$

where D is the diffusion constant for carriers (see Sec. 2.8) and η is a quantity definable by a transcendental equation

$$\eta \tan \eta = sd/2D$$

Where the condition $s < D/d$ is valid, the s - τ_s relation passes into an explicit relation:

$$\tau_s = d/2s \quad (2.43)$$

In practice, it is convenient to calculate the currents of excess carriers moving from the bulk to the surface disregarding the parameter τ_s , since the relation between the current density due to this carrier transport and the surface recombination velocity is rather simple:

$$i_s = qs\Delta n \quad (2.44)$$

Generally speaking, this current is parasitic and should be eliminated by decreasing the surface recombination rate with any means available.

2.7. Field Effect

The field effect is a *change in carrier concentration* (and, hence, in conductivity) of the surface layer of a semiconductor on exposure to an *electric field*. The layer having an increased concentration of majority carriers in comparison with that in the bulk is called *enriched*, and the layer with a decreased majority-carrier concentration is known as a *depletion* layer.

2.7.1. Nature of the field effect. Let voltage V be set up between a metal plate and semiconductor separated by a dielectric, air for example (Fig. 2.19). It is clear that in the metal-insulator-semiconductor (MIS) system, or MOS system, the flow of current is impossible. This system is thus in equilibrium and represents a peculiar capacitor in which one of the plates is a semiconductor. This plate will store the same amount of charge as the metal plate. But, in distinction to the charge on the metal plate, the charge in the semiconductor does not concentrate on its surface but spreads to a certain depth into the crystal bulk.

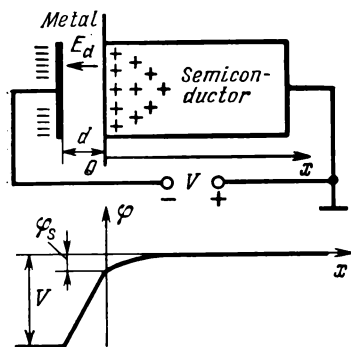


Fig. 2.19. Field effect in a MOS structure

An electric field produced by the voltage V is distributed between the dielectric and semiconductor. The field E_d in the dielectric is constant since space (bulk) charges in it are nonexistent, while the field in the semiconductor is certainly variable, because there is a space charge which decays in the direction away from the surface deep into the semiconductor.

The sign of charge in the semiconductor depends on the polarity of applied voltage. With negative polarity (Fig. 2.19), the charge induced in the semiconductor will be positive. In a p -type semiconductor, the positive charge is due to holes pulled up to the surface; in an n -type semiconductor, the positive charge stems from donor ions while they lose charge-compensating electrons.

The first case represents **enrichment** with and the second **depletion** of majority carriers in the surface layer. The reverse is true if we change the sign of applied voltage: in the n -type semiconductor the field will tend to enrich the surface with electrons, while in the p -type semiconductor the field will cause depletion of holes in the surface and "uncovering" of negative acceptor ions.

The distance over which **mobile** charges can exist in an enriched layer is called the *Debye length*, and the length of influence of **stationary** ion charges is called the *depletion layer depth*. The Debye

length is also defined as the *depth of penetration of an electric field* into a semiconductor, or, conversely, the *distance of shielding* of a semiconductor from an external electric field. We shall consider the Debye length and depletion layer depth in more detail in the further discussion. The enriched and depletion layers become thinner with an increase in the impurity concentration level and, hence, in the density of majority carriers. To put it in other words, thin layers are inherent in low-resistance semiconductors, and thick layers in high-resistance ones.

If we assume that a potential in the semiconductor bulk is equal to zero, the potential on the surface will be different from zero owing to the presence of charges between the bulk and surface. The difference in potentials between the surface and crystal bulk is termed the *surface potential* denoted by φ_s (see Fig. 2.19).

It should be noted that in the absence of an external field, the surface potential does not drop to zero, but has a finite **equilibrium** value, φ_{s0} , which is due to the presence of surface states that are able to capture or donate electrons for a comparatively long time (see p. 36). One more factor that influences the quantity φ_{s0} is a *contact potential difference between the metal and semiconductor* (see Subsec. 3.3.1). An external voltage required to compensate for the equilibrium surface voltage is referred to as the flat-band *voltage* and denoted as V_F .

We have mentioned earlier that the electric field spreads throughout the insulator and semiconductor. The field in the dielectric rises with a decreasing distance d until the dielectric breaks down. But even in a deep vacuum, where breakdown is impossible, the spacing d cannot be arbitrarily small, since at $d < 10$ nm the dielectric turns out to be permeable to mobile carriers on account of the *tunnel effect*—one of the phenomena attributed to the wave nature of electrons. The tunnel effect shows up in that an electron of certain potential energy overcomes a barrier of much higher potential by piercing the barrier if the thickness of the latter is sufficiently small. The probability of tunneling is defined by an exponent, $\exp(-10^8 d \sqrt{\Phi})$, where Φ is the barrier height, and d is the barrier thickness. This probability need be reckoned with at d in the order of 10 nm and below. As the tunnel effect becomes a reality, the MIS structure ceases to be an analog of the capacitor: carrier exchange through the insulator causes the flow of current and thus disturbs the equilibrium state. The current flow tends to decrease the charges on the "plates" until they disappear completely as the metal comes in electric contact with the semiconductor. Then the common conduction current begins to pass through the system.

2.7.2. General analysis. The distribution of a potential in the region of a space charge may in principle be evaluated with Pois-

son's one-dimensional equation:

$$\frac{d\varphi}{dx^2} = -\frac{\lambda}{\epsilon_0 \epsilon} \quad (2.45)$$

where λ is the density of charge, ϵ_0 is the electric constant (see Table 2.2), and ϵ is the relative permittivity of a semiconductor.

In the general case the charge density in a semiconductor is written in the form

$$\lambda = q(p + N_d^* - n - N_a^*) \quad (2.46)$$

where N_d^* and N_a^* are the concentrations of **ionized impurities**¹.

The concentration of free carriers on the right side of Eq. (2.46) is related to the electrostatic potential φ_E . To analyze this relation, we shall use Eqs. (2.11). Assume that in the semiconductor bulk, where the charges and field are absent, the electron and hole concentrations are equal to n_0 and p_0 , and the electrostatic potential is equal to φ_{E0} . Denote the corresponding quantities near the surface simply as n , p , and φ_E . Substitute into Eq. (2.11a) first the values of n_0 and φ_{E0} and then of n and φ_E , and equate the right sides (since in the equilibrium system $\varphi_E = \text{constant}$). We then get

$$\varphi_E - \varphi_{E0} = \varphi_T \ln(n_0/n)$$

Set for simplicity $\varphi_{E0} = 0$ (this condition corresponds to grounding of the semiconductor shown in Fig. 2.19). We can now express the concentration n in terms of φ_E :

$$n = n_0 \exp(-\varphi_E/\varphi_T) \quad (2.47a)$$

Using Eq. (2.11b), it is likewise possible to express p in terms of φ_E :

$$p = p_0 \exp(\varphi_E/\varphi_T) \quad (2.47b)$$

Change in Eqs. (2.47) the quantity φ_E for $-\varphi$ to pass from energy potential to electric potential (see p. 39). Substitute now the concentrations n and p into the right side of (2.46) and then the charge density λ into the Poisson equation. We obtain a nonlinear differential equation the analytic solution of which is generally nonexistent. But in two important individual cases, where there is a possibility of disregarding the ionized impurity concentration (enriched layers) or free carrier concentration (depletion layers), analytic solutions are existent. Consider these cases below.

2.7.3. Field effect in an intrinsic semiconductor. Substitute concentrations given by Eqs. (2.47) into the right-hand side of Eq. (2.46) and replace φ_E for $-\varphi$. Further, considering that the semiconductor is intrinsic, set $n_0 = p_0 = n_i$ and $N_d^* = N_a^* = 0$. We are now in

¹ In deriving neutrality condition (2.14), we have set $\lambda = 0$ and disregarded the concentration N_a^* because under consideration was an n -type semiconductor.

a position to reduce the charge density to the form

$$\lambda = -2qn_i \sinh (\varphi/\varphi_T)$$

Substitute this expression for λ into the right side of Eq. (2.45). Divide then both sides of the equation by φ_T and introduce the dimensionless variable $\Phi = \varphi/\varphi_T$. Last, the Poisson equation takes the form

$$\frac{d^2\Phi}{dx^2} = \frac{1}{l_{Di}^2} \sinh \Phi \quad (2.48)$$

where

$$l_{Di} = \sqrt{\frac{\epsilon_0 \epsilon \varphi_T}{2qn_i}} \quad (2.49)$$

is the Debye length in an intrinsic semiconductor. For silicon, $l_{Di} \approx 14 \mu\text{m}$.

Consider the simplest case where $|\varphi_s| < \varphi_T$, that is $|\Phi| < 1$. For the case in hand we may set $\sinh \Phi \approx \Phi$, and so Eq. (2.48) reduces to a linear differential first-order equation. For the boundary conditions, $\varphi(\infty) = 0$ and $\varphi(0) = \varphi_s$, the solution has the form

$$\varphi(x) \approx \varphi_s \exp(-x/l_{Di}) \quad (2.50)$$

From Eq. (2.50) it follows that the Debye length is the distance over which the potential drops off by a factor e as against the maximum value of φ_s on the surface.

Knowing the function $\varphi(x)$, it is easy to obtain the functions $E(x)$, $\lambda(x)$, $n(x)$, and $p(x)$. Fig. 2.20 illustrates all these functions plotted for the same polarity of voltage as that shown in Fig. 2.19. It also displays the band diagram of a semiconductor, where the curve $\varphi_E(x)$ and thus all other energy levels are mirror images of the curve $\varphi(x)$, as noted earlier on p. 39. Bending of energy bands near the semiconductor-insulator interface is a distinctive feature of the field effect.

On reversing the polarity of applied voltage, the bulk charge will change in sign and the bands will bend now in the opposite direction, downward. Irrespective of the polarity of voltage applied, the surface layer in an intrinsic semiconductor will be **rich** in carriers, either in electrons or in holes.

The surface potential may be found by relying on the continuity condition of electric induction at the interface between the semiconductor and dielectric:

$$\epsilon_s E(0) = \epsilon_d E_d(0) \quad (2.51a)$$

where ϵ_s and ϵ_d are the relative permittivities of the semiconductor and dielectric respectively.

The field in the dielectric is constant, and therefore (see Fig. 2.19)

$$E_d(0) = \frac{V - \varphi_s}{d} \quad (2.51b)$$

The field in the semiconductor at the boundary of the dielectric is determined by the function $\varphi(x)$:

$$E(0) = -\frac{d\varphi}{dx}(0) \quad (2.51c)$$

Omitting mathematic calculations, we represent the function $\varphi_s(V)$ in the form of curves shown in Fig. 2.21. From these curves it is seen that the surface potential gains a greater portion of applied voltage

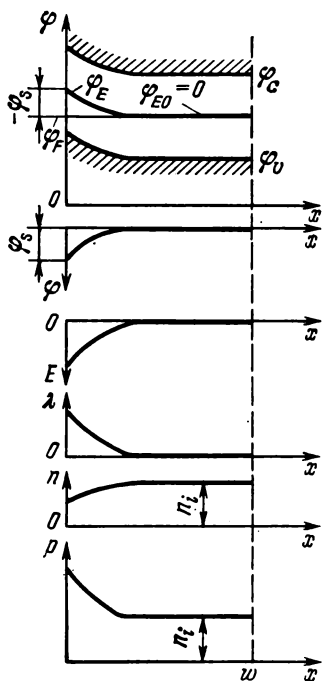
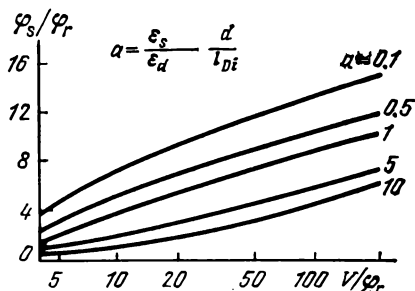


Fig. 2.20. Field effect in an intrinsic semiconductor; band diagram and distribution of a potential, charge, and carrier concentrations

Fig. 2.21. Surface potential in an intrinsic semiconductor as a function of insulator thickness and voltage on a metal electrode



with a decrease in the dielectric thickness (in the parameter a). At any real values of the dielectric thickness and applied voltage, the surface potential does not exceed a few tenths of a volt.

2.7.4. Field effect in extrinsic semiconductors. What distinguishes the field effect in **impurity** semiconductors from that in pure semiconductors is the possibility of producing both enriched and depletion layers.

The *enhancement mode* sets in where the polarity of applied voltage is such that the electric field **pulls** majority carriers to the surface. This case resembles the one shown in Fig. 2.20, but differs from the latter in a lesser amount of band bending (Fig. 2.22a). A smaller bend of band edges is due to the fact that an impurity semiconductor is rich in mobile carriers, so even a minor surface potential is enough to ensure the required charge near the surface.

Given the condition $\varphi_s < 2\varphi_T$, the potential in an extrinsic semiconductor is described by Eq. (2.50), but the Debye length has the form

$$l_D = \sqrt{\frac{\epsilon_0 \epsilon \varphi_T}{qN}} \quad (2.52)$$

where N is the concentration of an ionized impurity (either donor or acceptor).

Since $N \gg n_i$, the Debye length in impurity semiconductors is much smaller than in intrinsic semiconductors. Besides, it does not

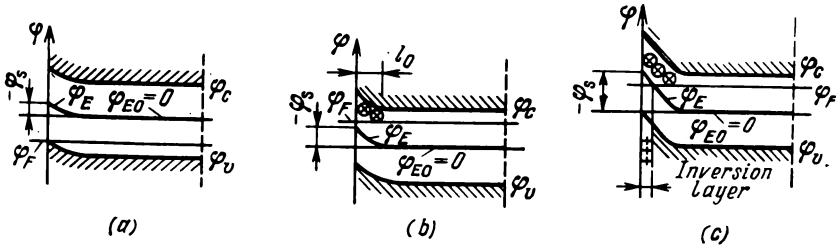


Fig. 2.22. Field effect in extrinsic semiconductors

(a) enhancement mode; (b) depletion mode; (c) formation of an inversion layer

practically depend on the material. Assuming $N = 10^{16} \text{ cm}^{-3}$, from Eq. (2.52) we find the typical value of l_D which is equal to about $0.04 \mu\text{m}$. As seen, the field penetrates extrinsic semiconductors to an insignificant depth.

If we apply (2.52) to metals, though this is not totally warrantable, we can find that with huge free carrier concentrations (10^{22} to 10^{23} cm^{-3}) inherent in metals, the Debye length l_D spans a range merely within tenths of a nanometer, which is equal to 1 or 2 atomic spacings. This estimate is a good illustration of the fact that charges in a metal always gather on the surface, and *both charges and an electric field are not found inside the metal*.

At a sufficiently large voltage, the surface potential rises so heavily that the Fermi level in the surface region will be found to lie within one of the energy bands (in the valence band in Fig. 2.22a). In this region the semiconductor degenerates and converts to a semimetal. The boundary portion of the MOS system thus changes to a metal-

insulator (oxide)-semimetal system which represents an **ordinary** capacitor; the main part of the semiconductor is equivalent to a resistor connected in series with the capacitor. Since the voltage drop in the semimetallic layer is negligible, the surface potential does not vary any longer and remains close in value to φ_{sm} at which the semimetal has formed.

The *depletion mode* sets in under such a polarity of applied voltage that causes the majority carriers **to repel** from the surface. In this case the surface potential can be much larger in value than when the enhancement process takes place (Fig. 2.22*b*). As mentioned earlier, repulsion of majority carriers leads to the appearance of an uncompensated space charge of impurity ions.

Suppose the depletion layer has a **sharp** boundary a distance l_0 from the surface. Set the space charge density in the depletion layer constant and equal to qN , where N is the ionized impurity concentration. Substituting the quantity $\lambda = qN$ in Poisson's equation (2.45) and using the boundary values, $E(l_0) = 0$ and $\varphi(l_0) = 0$, we get after double integration:

$$\varphi = \frac{\lambda}{2\epsilon_0 e} (x - l_0)^2$$

Setting $x = 0$ and $\varphi(0) = \varphi_s$, from this equation we find the length (thickness) of the depletion layer:

$$l_0 = \sqrt{\frac{2\epsilon_0 e |\varphi_s|}{qN}} \quad (2.53)$$

Though the structure of Eqs. (2.53) and (2.52) is the same, there is a substantial difference between these expressions: the Debye length **only depends on the properties of a material**, while the thickness of the space charge also depends on the **applied voltage** since it determines the potential φ_s (see Fig. 2.21). The quantity l_0 is commonly a few times as large as the quantity l_D .

As the voltage rises, the field continues sweeping away majority carriers (causing the depletion layer to widen), but at the same time **minority** carriers are being pulled over to the surface. When the minority carrier charge exceeds the charge of the remaining majority carriers, *the conductivity type of the surface layer changes*. This phenomenon is known as *inversion of the conductivity type*, and the layer formed by minority carriers as *inversion layer* (Fig. 2.22*c*).

From the standpoint of band theory, the formation of an inversion layer is due to the fact that near the surface **the electrostatic potential level crosses the Fermi level**. In the surface layer, therefore, the Fermi level shifts to that half of the band-gap where minority carriers prevail. The thickness of an inversion layer comes to merely 1 or 2 nm, or 3 or 4 lattice constants.

From Fig. 2.22c it is obvious that the inversion layer forms at a surface potential $-(\varphi_F - \varphi_{E0})$. A yet higher rise in external voltage entails a further increase of the potential φ_s until the Fermi level crosses over the boundary of an energy band (valence band in Fig. 2.22c). The boundary layer then turns to a semimetal, while the potential φ_s practically stops to vary (see p. 64) and remains equal to

$$\varphi_{sm} = -2(\varphi_F - \varphi_{E0}) \quad (2.54)$$

The maximum value of surface potential commonly ranges from 0.6 to 1.0 V.

2.8. Behavior of Carriers in Semiconductors

In the general case, charge carrier transport results from two processes: diffusion due to the concentration gradient and drift due to the electric potential gradient. Since there are two types of carrier, electrons and holes, the total current comprises four components:

$$j = (j_n)_{dr} + (j_n)_{dif} + (j_p)_{dr} + (j_p)_{dif} \quad (2.55)$$

where subscripts "dr" and "dif" relate to drift and diffusion components respectively.

In the analysis of transistor behavior, it is more convenient to use current densities j , as we have done in Eq. (2.55), rather than currents. However, for brevity we shall call the quantity j just current where this term cannot cause misunderstanding.

2.8.1. Current components. For the one-dimensional case¹, where the motion of carriers occurs only along the x -axis, the electron and hole drift current densities are given by

$$(j_n)_{dr} = qn\mu_n E = -qn\mu_n (\partial\varphi/\partial x) \quad (2.56a)$$

$$(j_p)_{dr} = qp\mu_p E = -qp\mu_p (\partial\varphi/\partial x) \quad (2.56b)$$

For diffusion current densities, one must use chemical potential gradients of respective carriers instead of electric potential gradients. Chemical potentials are the second terms on the right of Eqs. (2.11). Differentiate these terms with respect to x and substitute into expressions (2.56) the found values for E . The diffusion currents will then be described by

$$(j_n)_{dif} = q\mu_n \varphi_T \frac{dn}{dx} = qD_n \frac{dn}{dx} \quad (2.57a)$$

for electrons, and by

$$(j_p)_{dif} = -q\mu_p \varphi_T \frac{dp}{dx} = -qD_p \frac{dp}{dx} \quad (2.57b)$$

¹ The subsections that follow will deal only with one-dimensional processes.

for holes, where D_n and D_p are the *diffusion constants* for electrons and holes respectively. These quantities play the same role in the diffusion process as the mobilities in the drift process. The diffusion constants and mobilities are connected by the Einstein relationship:

$$D = \varphi_T \mu \quad (2.58)$$

The values of diffusion constants are given in Table 2.1.

From the comparison of Eqs. (2.56) with Eqs. (2.57), we can infer that drift current components are proportional to carrier concentrations, whereas diffusion current components are independent of carrier concentrations and are only the functions of **concentration gradients**.

2.8.2. Continuity equations. Expressions (2.56) and (2.57) say that for the estimation of total current [see Eq. (2.55)], we must know concentration distribution functions $n(x)$ and $p(x)$, apart from the potential distribution function $\varphi(x)$.

In the general case, concentrations depend not only on the coordinates but also on time; in other words, we have to deal with the functions of two variables: $n(x, t)$ and $p(x, t)$. These functions are the solutions of the so-called *continuity equations for carrier flows*. For electron and hole densities, the continuity equations have the form

$$\frac{dn}{dt} = \Delta g - \frac{n - n_0}{\tau} + \frac{1}{q} \operatorname{div}(j_n) \quad (2.59a)$$

$$\frac{dp}{dt} = \Delta g - \frac{p - p_0}{\tau} - \frac{1}{q} \operatorname{div}(j_p) \quad (2.59b)$$

where $n - n_0 = \Delta n$ and $p - p_0 = \Delta p$ are excess concentrations [see Eqs. (2.28)], Δg is the rate of carrier generation by **external** influences such as light, and τ is the lifetime of excess carriers.

It is easy to see that the continuity equations generalize the carrier accumulation equation (2.33) derived in the analysis of recombination processes. The extension of (2.33) involves the inclusion in its right-hand side of one more factor that changes the concentration in the presence of current, namely, flow vector divergence j/q , along with generation and recombination terms.

For the one-dimensional case

$$\operatorname{div}\left(\frac{j}{q}\right) = \frac{1}{q} \frac{\partial}{\partial x} j$$

If we perform this operation for all four current components described by Eqs. (2.56) and (2.57), substitute the found values into (2.59), and omit the generation term Δg , the continuity equations will take

on the form

$$\frac{\partial n}{\partial t} = -\frac{n-n_0}{\tau} + D_n \frac{\partial^2 n}{\partial x^2} + \mu_n E \frac{\partial n}{\partial x} + n\mu_n \frac{\partial E}{\partial x} \quad (2.60a)$$

$$\frac{\partial p}{\partial t} = -\frac{p-p_0}{\tau} + D_p \frac{\partial^2 p}{\partial x^2} - \mu_p E \frac{\partial p}{\partial x} - p\mu_p \frac{\partial E}{\partial x} \quad (2.60b)$$

The last terms on the right of Eqs. (2.60) allow for **space charges present** in a semiconductor. It should be noted in passing that in the presence of space charges, one has to make use of Poisson's equation (2.45) along with continuity equations, in performing the analysis. Under the conditions of neutrality the last terms are absent. The third terms are independent of bulk charges; the inclusion of these terms is necessary in a number of cases, for instance, where semiconductors exhibit a **built-in** electric field, which is typical of inhomogeneous semiconductors (see p. 42).

If the electric field is absent or its influence is negligibly small, we can set $E = 0$ and thus simplify the general continuity equations:

$$\frac{\partial n}{\partial t} = -\frac{n-n_0}{\tau} + D_n \frac{\partial^2 n}{\partial x^2} \quad (2.61a)$$

$$\frac{\partial p}{\partial t} = -\frac{p-p_0}{\tau} + D_p \frac{\partial^2 p}{\partial x^2} \quad (2.61b)$$

These are *diffusion equations* which find wide use in the analysis of semiconductor devices.

2.8.3. Carrier diffusion. Let a scattered beam of light fall on the surface of a semiconductor (Fig. 2.23). The light penetrating a thin surface layer causes generation of electron-hole pairs in the region of influence. This gives rise to electron and hole concentration gradients between the surface and crystal bulk, so that excess carriers will start diffusing into the bulk of the semiconductor. Such a joint motion of both types of carrier is called *bipolar* or *ambipolar* diffusion.

If the mobilities and thus diffusion constants for electrons and holes were the same, the carriers of both types would move as a **single neutral flow**. In reality carriers differ in mobility, therefore the electron flow tends to outrun the hole flow. A slight mutual shift of the flows results in a small space charge and an electric field which impedes the electron flow and accelerates the hole flow. In the end, the process comes to a **steady state**: excess

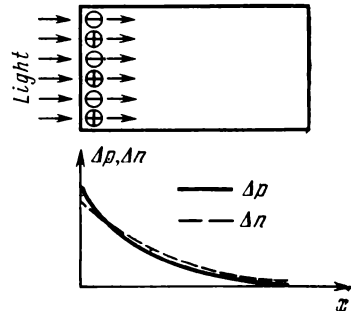


Fig. 2.23. Ambipolar diffusion; Dember effect

electrons and holes gather in "clouds" shifted one relative to the other. These clouds move in step, and so the **resultant current is absent**. The carrier density in clouds decreases in the direction away from the surface on account of recombination.

The described phenomenon is known as the *Dember effect*, and the electric field and potential difference typical of this effect as the *Dember field* and *Dember voltage*.

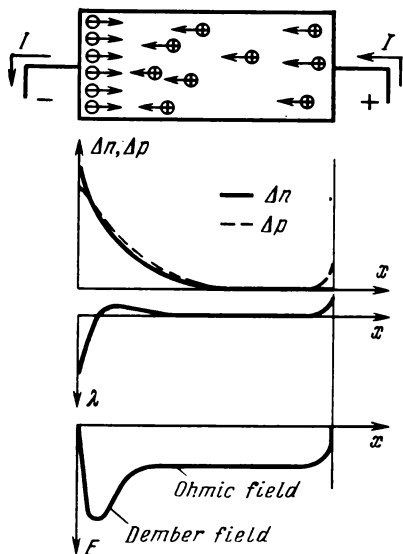


Fig. 2.24. Monopolar diffusion; carrier injection

This effect is rather strong only at large excess densities and large resistivities of semiconductors.

It is *monopolar diffusion* that finds most extensive practical use¹. Characteristic of monopolar diffusion is the enrichment of the surface layer with carriers of only **one** type, **minority** carriers (Fig. 2.24). The process of minority carrier introduction into the surface layer by an appropriate method is known as *injection*.

Assume for clarity that the injection process involves the introduction of electrons into a *p*-type semiconductor. The injected electrons will diffuse into the crystal bulk owing to the gradient in concentration, and thus the flow of electrons will appear in the semiconductor. The charge of excess electrons will practically be balanced

at an instant (for the dielectric relaxation time, see p. 44) by the equivalent charge of holes pulled from deep layers. As a result, near the surface of injection there appears a quasineutral electron-hole cloud almost identical to the cloud produced in bipolar diffusion. Despite this formal similarity of clouds, monopolar diffusion principally differs from bipolar diffusion by the following features.

1. The presence of current suggests that the semiconductor is an element of a closed circuit; hence, along with the Dember field concentrated near the injection surface, there is an ordinary ohmic field in the crystal bulk set up by the applied voltage (see Fig. 2.24).

2. Electrons and holes travel in **opposite** directions: electrons move into the crystal bulk, while holes head for the injection surface to reach the region of the electron-hole cloud, where intensive recombina-

¹ Since monopolar diffusion is a widespread phenomenon, it is common to omit the modifier "monopolar".

nation takes place and there is a need for replenishment with new majority carriers.

3. The total current keeps constant, while its electron and hole components vary in opposite directions: farther away from the surface, the electron current decays because of recombination, whereas the hole current grows. That is why the hole component plays the main role rather far from the surface and displays a purely drift character, since the holes travel in the field produced by the external voltage. On the contrary, in the immediate vicinity to the surface, the current is almost purely electronic which results from the **diffusion** process, because the field strength here is close to zero (see Fig. 2.24).

The problem in the distribution of carriers during the diffusion process is difficult to solve accurately. This problem is usually solved in a "*diffusion*" *approximation* for low excess concentrations, or what is said low injection levels.

The diffusion approximation implies the use of diffusion equations (2.61), though, as is known, the electric field is present in a semiconductor. In each concrete case one must assess whether or not the diffusion approximation is valid.

The *injection level* is the ratio of the excess carrier concentration to the equilibrium **majority**-carrier concentration, or, which is the same, to the impurity concentration:

$$\delta = \Delta n / N \quad (2.62)$$

A **low** injection level is considered to be the level whose value is much smaller than unity. In this case the inequality

$$\Delta n \ll n_0 + p_0 \quad (2.63)$$

holds true. We have used this inequality earlier in deriving Eqs. (2.32) and (2.38). The condition for a low injection level ensures the **linearity** of diffusion equations. Under the conditions of neutrality, Δn is approximately equal to Δp , therefore Eqs. (2.62) and (2.63) are valid for both types of carrier.

2.8.4. Analysis of diffusion processes. If only **excess** carriers are of interest, as is indeed so in most cases, then it is enough to use **one** of the two diffusion equations, since the other gives the same result on account of the neutrality condition, $\Delta n \approx \Delta p$. In reality the functions $\Delta n(x)$ and $\Delta p(x)$ differ somewhat because of the difference between the diffusion constants D_n and D_p , that is, as a result of the **Dember** effect. The diffusion approximation, however, disregards electric fields, including the **Dember** field.

Find the excess concentration from Eq. (2.61a). For this, substitute $n = n_0 + \Delta n$ and omit the subscript n in the diffusion constant. Next, divide both sides of the equation by D . The diffusion equation

will then take the form

$$\frac{\partial^2 (\Delta n)}{\partial x^2} - \frac{\Delta n}{L^2} = \frac{1}{D} \frac{\partial (\Delta n)}{\partial t} \quad (2.64)$$

A steady-state variant of the equation results if we set $\partial (\Delta n)/\partial t$ equal to zero:

$$\frac{d^2 (\Delta n)}{dx^2} - \frac{\Delta n}{L^2} = 0 \quad (2.65)$$

The quantity L entering Eqs. (2.64) and (2.65) is given by

$$L = \sqrt{D\tau} \quad (2.66)$$

This quantity is a *mean diffusion length* which defines the average distance to which the minority carriers can diffuse during their

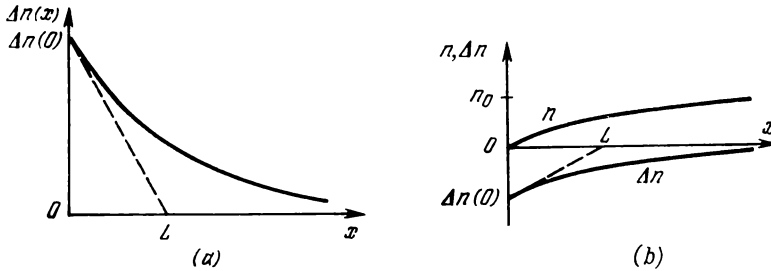


Fig. 2.25. Stationary distribution of excess carriers in diffusion (a) and extraction (b)

lifetime. The ratio L/τ is the *mean rate of carrier diffusion*. The diffusion length is one of the fundamental quantities in semiconductor physics and technology. The typical values of L for silicon range from 5 to 20 μm depending on the life-time.

The steady-state equation (2.65) is an ordinary linear second-order equation. Its solution represents a sum of the exponents:

$$\Delta n(x) = A_1 \exp(x/L) + A_2 \exp(-x/L) \quad (2.67)$$

where coefficients A_1 and A_2 are determined from boundary conditions. Assume $\Delta n(\infty) = 0$, in other words, suppose that in a semiconductor portion, a certain distance away from the injection surface, the excess concentrations are absent and this portion of semiconductors is in equilibrium. At such a boundary condition, $A_1 = 0$. Setting $x = 0$, we get $A_2 = \Delta n(0)$; consequently, the distribution of excess concentration takes an exponential form (Fig. 2.25a):

$$\Delta n(x) = \Delta n(0) \exp(-x/L) \quad (2.68)$$

From this expression and Fig. 2.25a it follows that within the diffusion length the excess concentration decreases by a factor of e . In

the region $3L$ or $4L$ long the concentration drops off by a factor of 20 to 50 and thus becomes negligible in comparison with the boundary concentration.

By differentiating Eq. (2.68), we obtain the gradient in concentration:

$$\frac{d(\Delta n)}{dx} = -\frac{\Delta n(0)}{L} \exp(-x/L) \quad (2.69a)$$

As seen, the concentration gradient and thus **the diffusion current decays in the direction away from the surface** into the depth of the semiconductor. The gradient is at a maximum at $x = 0$, that is, on the surface of injection:

$$\left. \frac{d(\Delta n)}{dx} \right|_{x=0} = -\frac{\Delta n(0)}{L} \quad (2.69b)$$

The **transient** equation (2.64) is a linear differential equation of the second order in partial derivatives. A few methods are applicable for its solution. In technical practice the Laplacian operator method is most popular.

With the operator method, the time function, which is $\Delta n(x, t)$ in our case of interest, is replaced by the Laplace transform $\Delta n(x, s)$, and the time derivative by the quantity

$$s[\Delta n(x, s) - \Delta n(x)_{t=0}] \quad (2.70)$$

where s is the Laplacian operator¹. In solving an operational equation, **we take the operator as an algebraic factor** and find the transform of the sought-for function $\Delta n(x, t)$. This function (the *original* function corresponding to the transform) can in most cases be obtained from special tables.

Assume that initially a semiconductor stays in equilibrium. Based on this assumption, we can set $\Delta n(x) = 0$ in (2.70). Changing the derivative $\partial(\Delta n)/\partial t$ in the right side of (2.64) for its transform $s\Delta n$ gives an ordinary differential equation

$$\frac{d^2(\Delta n)}{dx^2} - \frac{1}{L^2} \Delta n = \frac{s}{D} \Delta n$$

Next multiply and divide its right side by τ , replace the product $D\tau$ by L^2 in accordance with Eq. (2.66), and combine the terms with Δn . As a result we derive a differential equation in the operator form

$$\frac{d^2(\Delta n)}{dx^2} - \frac{1+s\tau}{L^2} \Delta n = 0 \quad (2.71a)$$

This same expression can be written in a more illustrative form:

$$\frac{d^2(\Delta n)}{dx^2} - \frac{\Delta n}{L^2(s)} = 0 \quad (2.71b)$$

¹ Here and elsewhere in this book we shall designate the Laplacian as s rather than as the commonly accepted symbol p to avoid confusion with the designation of hole density.

where $L(s)$ is the operational diffusion length:

$$L(s) = \frac{L}{\sqrt{1+s\tau}} \quad (2.72)$$

Since the form of Eq. (2.71b) is the same as that of Eq. (2.65), the solution of both must formally coincide:

$$\Delta n(x, s) = \Delta n(0) \exp[-x/L(s)] \quad (2.73)$$

This is the Laplace transform of the sought-for function $\Delta n(x, t)$. The function itself (the original function of this transform) need be found from correspondence (operator) tables.

The diffusion length in operator form is merely a symbol of the mathematic operation, inseparable from Eqs. (2.71). Therefore,

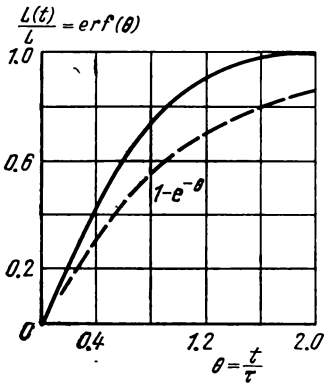


Fig. 2.26. Error function and its approximation (dash line)

strictly speaking, the original time function, $L(t)$, cannot be employed as an independent quantity. Nevertheless this function proves useful for purely qualitative estimations.

From operator tables it follows that the original time function of the transform given by (2.72) takes the form

$$L(t) = L \text{erf}(t/\tau) \quad (2.74)$$

where $\text{erf}(t/\tau)$ is the error function expressed as

$$\text{erf}(\theta) = \frac{2}{\sqrt{\pi}} \int_0^\theta e^{-\xi^2} d\xi$$

Its derivative has the form $(2/\sqrt{\pi})e^{-\theta^2}$. A **complementary error function** is the function $\text{erfc}(\theta) = 1 - \text{erf}(\theta)$. Its derivative differs only in sign. The graph of the error function appears in Fig. 2.26. For comparison, the figure also illustrates the elementary function $1 - e^{-\theta}$, shown as a dash line, which resembles the error function

in shape but is less steep. It is safe to say that the diffusion length increases approximately in an exponential manner with a time constant somewhat smaller than τ .

The qualitative conclusions relating to the transient process are the following.

At the start, when $L(0) = 0$, the concentration gradient near the surface of injection, according to Eq. (2.69b), proves infinite and, hence, excess carriers diffuse in the crystal at a great velocity. As the quantity $L(t)$ rises, the concentration gradient of the surface gets

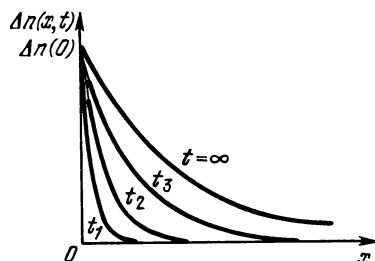


Fig. 2.27. Distribution of injected carriers during the transient

lower and the diffusion rate progressively decreases. Finally, at $t \approx 2\tau$, the steady state sets in. Based on these considerations, it has been possible to plot typical concentration distribution curves for a few moments of the transient (Fig. 2.27).

To this point the excess concentration Δn has been considered positive since injection results in **additional** carriers. However, the reverse process is possible, in which a part of equilibrium carriers **are drawn** from the surface layer. Such a process is called *extraction*.

The excess concentration in the case of extraction will obviously be negative (see Fig. 2.25b) because the quantity of carriers decreases in comparison with that typical for the equilibrium state. Besides the process of extraction gives rise to the concentration gradient of the other sign than would be found in the injection process since the flow of minority carriers is directed to the surface rather than into the crystal bulk. An important feature of extraction as against injection is the fact that the excess concentration cannot exceed the equilibrium concentration of minority carriers, n_0 in Fig. 2.25b.

2.8.5. Combined motion of carriers. If the injection of carriers occurs in an **inhomogeneous** semiconductor which has an internal electric field (see Subsec. 2.4.7) the diffusion of carriers will add to carrier drift to effect a *combined* carrier motion. For the analysis of this case, we would have, generally speaking, to utilize continuity equations (2.60), which presents considerable difficulties. But since in practice

one commonly has to deal with a **uniform** field $E = \text{constant}$, the problem becomes much simpler.

In most inhomogeneous semiconductors employed in microelectronics, the law of impurity distribution approaches the exponential function (see Sec. 6.5):

$$N(x) = N(0) e^{-x/L_N} \quad (2.75)$$

Here the coordinate x is counted off from the surface. The quantity L_N is a *mean depth of doping* (at a distance L_N from the surface, the impurity concentration decreases by a factor of e).

Assuming that majority carriers, holes for example, obey the same law of distribution as impurity atoms, we get

$$p(x) = p(0) e^{-x/L_N} \quad (2.76)$$

In Subsec. 2.4.7 we have given the expression for the internal field strength in an inhomogeneous p -type semiconductor:

$$E = \varphi_T \frac{dp/dx}{p}$$

Substitute the concentration given by Eq. (2.76) and its derivative in this expression. The field strength expression then takes the form

$$E = -\varphi_T/L_N \quad (2.77)$$

Thus, *in semiconductors exhibiting the exponential distribution of impurities, the electric field is uniform*, $E = \text{constant}$. We shall rely on this deduction in analyzing the combined motion of carriers.

Since space charges in the region of a uniform field are absent, we should set $\partial E/\partial x = 0$ in Eqs. (2.60). For the same reason we can regard the semiconductor under analysis as being neutral and thus assume that excess concentrations Δn and Δp are equal. In this connection, it becomes possible to use **one** of the continuity equations, as we have done earlier. Assuming the semiconductor is the p -type, take Eq. (2.60a) for minority carriers, electrons. Last, we restrict ourselves to **steady-state** conditions and thus put $\partial n/\partial t = 0$. Keeping in mind the above stipulations, divide both sides of Eq. (2.60a) by D having regard to Eq. (2.58). The expression then reduces to the form

$$\frac{d^2(\Delta n)}{dx^2} + \frac{E}{\varphi_T} \frac{d(\Delta n)}{dx} - \frac{\Delta n}{L^2} = 0$$

Introduce the dimensionless *field-form factor* that defines the field strength E :

$$\theta = \frac{E}{2\varphi_T/L} \quad (2.78)$$

Using this factor, we obtain the differential equation of the form

$$\frac{d^2(\Delta n)}{dx^2} + 2\frac{\theta}{L}\frac{d(\Delta n)}{dx} - \frac{\Delta n}{L^2} = 0 \quad (2.79)$$

The solution to Eq. (2.79) will be the same in form as Eq. (2.68):

$$\Delta n(x) = \Delta n(0) e^{-x/\mathcal{L}} \quad (2.80)$$

but the diffusion length here is replaced by the quantity

$$\mathcal{L} = \frac{L}{\sqrt{\theta^2 + 1} + \theta} \quad (2.81)$$

Formula (2.81) is derived for a *p*-type semiconductor in which the *minority* carriers are electrons. For an *n*-type semiconductor whose minority carriers are holes, a minus sign in front of θ must stand in the denominator of Eq. (2.81).

The quantity \mathcal{L} is called the *depth of pulling*. In the combined motion of carriers, this quantity plays the same part as the diffusion length in purely diffusion carrier transport. But these two parameters differ in value; namely, in an **accelerating** field (for electrons this means that $E < 0$ and $\theta < 0$), $\mathcal{L} > L$, and thus carriers penetrate deeper into the crystal than in the absence of the field. In a **brake** field, on the contrary, $\mathcal{L} < L$, and carriers move to a *smaller* depth.

For almost purely diffusion motion, θ must be smaller than 0.2 or 0.3, and for almost purely drift motion, θ must be larger than 2 or 3.

3.1. General

Homogeneous semiconductors and semiconductor layers find rather limited uses, mainly as resistors of various types. The basic integrated elements and the major range of discrete semiconductor devices generally have inhomogeneous structures. Two important variants of these structures are a *pn* junction (contact between two semiconductors of different conductivity types) and an MS structure (metal-semiconductor contact, or junction).

This chapter considers most thoroughly *pn* junctions since they form the basis of modern microelectronics, and also MS junctions capable of serving the functions of both diodes and conventional ohmic contacts. The latter are inevitably present in any semiconductor device and IC. The last section of this Chapter covers the phenomena that appear at the contact between a semiconductor and an insulator, primarily between silicon and silica (silicon dioxide). These phenomena exert an influence on the characteristics of *pn* junctions and especially do so when using the field effect.

3.2. Electron-Hole Junctions

A combination of two semiconductor layers with different types of conductivity (Fig. 3.1a) exhibits rectifying or gating properties: such a structure allows the current to pass through more readily in

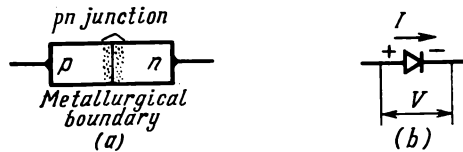


Fig. 3.1. Semiconductor diode

(a) simplified structure; (b) graphical symbol of diode

one direction than in the other. The polarity of voltage at which large currents flow through the system is called forward, and the opposite polarity that causes the flow of small currents is termed reverse. The common terms for voltages and currents here are *forward* and *reverse (bias) voltages* and *forward* and *reverse currents*.

Owing to its rectifying property, the structure being considered can serve as a semiconductor diode. Fig. 3.1b illustrates the graphi-

cal symbol for a diode with the direction of forward current and the polarity of forward voltage.

The surface over which the p - and n -layer come in contact is called a *metallurgical boundary*, and the adjacent region of space charges is known as an electron-hole junction, or *pn junction*¹. The other two (external) contacts in the diode do not exhibit rectification properties, and therefore are called ohmic.

3.2.1. Structure of a pn junction. Electron-hole junctions fall into two large classes according to the degree of abruptness of the metallurgical boundary and the relation between the resistivities of n -type and p -type layers.

An *abrupt (step) junction* is the junction with an ideal boundary having on one side donors and on the other acceptors of constant concentrations, N_d and N_a . Such junctions are the simplest to deal with in the analysis, and therefore it is usual practice to consider all real junctions as abrupt, where possible.

A *graded junction* is the junction in which the concentration of one type of impurity in the region of the boundary decreases and that of the other type grows. The boundary itself lies where the impurity concentrations are equal ($N_d = N_a$), that is, where the semiconductor is found to be compensated (see Subsec. 2.4.2.). All real pn junctions are graded, the degree of grading being the function of the effective concentration gradient in the region of the pn boundary.

The rectifying property is only evident in pn junctions whose boundary region has an effective concentration gradient that satisfies the inequality

$$\frac{dN}{dx} \gg \frac{n_i}{l_{Di}}$$

where N is the effective concentration of an impurity (see p. 36), and l_{Di} is the Debye length in an intrinsic semiconductor [see Eq. (2.49)]. For silicon the required value of the concentration gradient is $dN/dx \gg 10^{13} \text{ cm}^{-4}$.

Depending on the impurity concentration levels in p - and n -layers, there are *symmetric*, *asymmetric*, and *one-sided* junctions. In symmetric junctions, impurity concentrations N_{dn} and N_{ap} in respective layers are approximately equal. Symmetric junctions are not typical for semiconductor engineering. In wide use are mainly asymmetric junctions in which impurity concentrations are different.

One-sided junctions represent the case of sharp asymmetry, where impurity concentrations and thus majority carrier densities differ by a factor of 10 or 10^2 or even more. These junctions are designated

¹ The explanation of why space charges are present near the metallurgical boundary will be given below.

as n^+p or p^+n , where the upper index “+” stands for the layer with a much higher concentration.

In the text below, we shall use rarely the index “+”, but imply that the pn junction under discussion is one-sided or at least asymmetric.

In Fig. 3.2 is shown the electrical model of a pn junction and its origin. For the illustrative purpose, the difference in concentrations between majority carriers, n_{n0} and p_{p0} , is taken to be smaller than it is in reality.

Since the electron concentration in an n layer is much greater than in a p layer, a part of electrons diffuse from the n - into the p -layer, and so excess electrons appear in the p layer near the pn interface. These electrons recombine with holes until the equilibrium condition as given by Eq. (2.9) is met. Since the hole concentration in

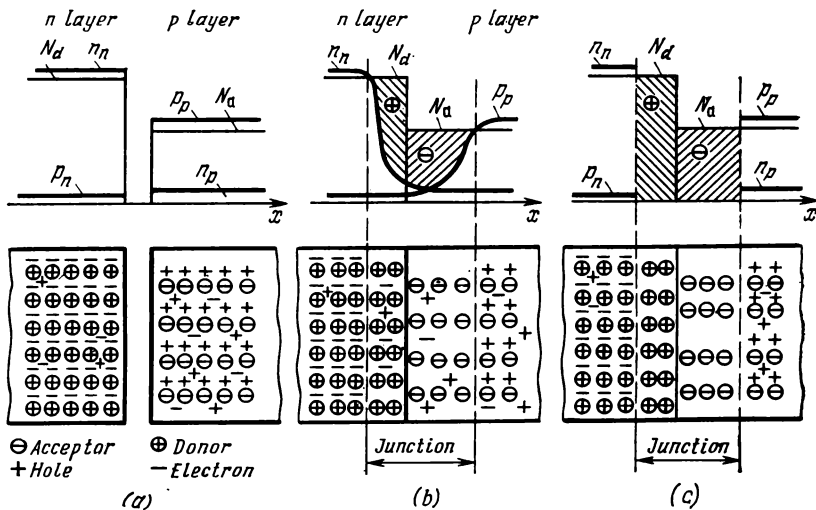


Fig. 3.2. Electrical model of a pn junction

(a) initial state of layers; (b) space charges in real junction; (c) space charges in ideal junction

this region decreases, uncompensated negative charges of acceptor atoms are left uncovered. On the left of the boundary, the diffusion of electrons leaves uncovered uncompensated positive charges of donor atoms (Fig. 3.2b). A similar reasoning may apply to holes, which diffuse from the p layer into the n layer. However, in a **one-sided** junction, where $p_{p0} \ll n_{n0}$, the transport of holes is insignificant because their concentration gradient is substantially lower than that of electrons.

The space charges so built up and the attendant fields ensure the Boltzmann equilibrium in the pn junction region. The region of

space charges is called a *depletion layer* because both portions of this region have a sharply decreased concentration of mobile carriers.

In most cases, it is possible to idealize a *pn* junction as is shown in Fig. 3.2c, that is, to neglect **completely** the presence of free carriers at the junction and regard the junction boundaries as ideally abrupt. This approach simplifies the solution of problems.

The junction on the whole is neutral: the positive charge on the left is equal to the negative charge on the right of the metallurgical boundary. The charge **densities**, however, are sharply different because of the difference in impurity concentrations. For this reason the **lengths** (widths) of depletion layers are different too: in the layer of a lower impurity concentration (*p* layer in our case), the space charge region is noticeably wider. The *asymmetric junction* is said to lie in a *high-resistance layer*.

Figure 3.3 illustrates the carrier distribution in a semilog scale which is more convenient for quantitative estimation and comparison. One should pay attention to the fact that inside the *pn* junction

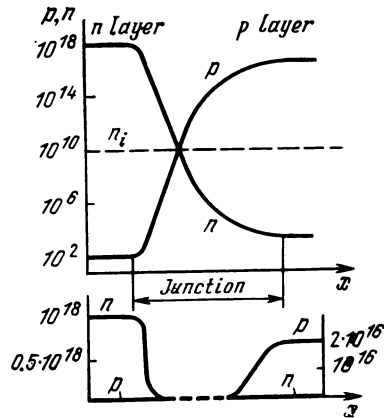


Fig. 3.3. Carrier distribution in an asymmetric junction (semilog and linear scales)

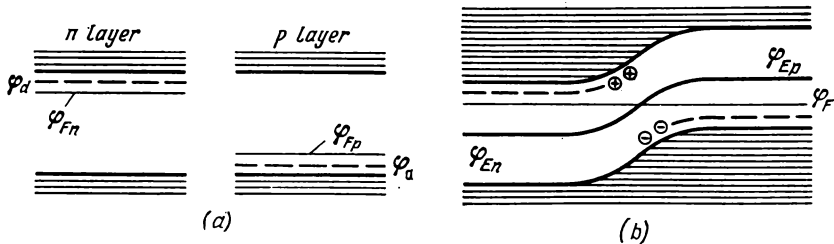


Fig. 3.4. Band diagrams of layers (a) and a *pn* junction in equilibrium state (b)

there is a region with an intrinsic (minimum) concentration of carriers. That is why the *junction region shows the highest resistance in comparison with the rest of the diode structure*. The resistivity in this region is a few orders of magnitude higher than the resistivity of neutral *n*-type and *p*-type regions.

Figure 3.4 shows the band diagrams of a *pn* junction before and after imaginary "contact" between the layers. As seen, because

the Fermi level remains constant throughout in the equilibrium *pn* junction, the band edges bend in the boundary region. This results in a potential difference (potential barrier) and an electric field typical of the Boltzmann equilibrium condition.

To characterize the equilibrium state of the junction, it is possible to resort to a descriptive interpretation, that is, to liken the electrons to heavy balls capable of moving along the bottom of the conduction band. The band model then appears to be like that in Fig. 3.5.

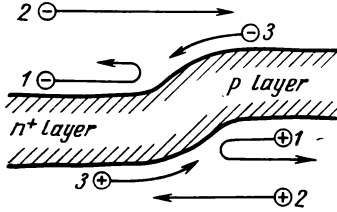


Fig. 3.5. Band model of carrier motion at the junction

Most of the electrons 1 in the n^+ layer have a small thermal energy. Such electrons “roll” along the conduction band bottom toward the barrier but cannot overcome it; having penetrated the junction to a certain depth, these electrons “rebound” and come back to the n^+ layer. And only electrons 2 have enough energy to jump over the barrier and get to the p layer, thus forming the carrier flow heading from left to right. As for electrons 3 in the p layer, these *roll down unimpeded* to the n^+ layer irrespective of the energy, forming the counterflow from right to left. These two flows of electrons get into equilibrium. A similar model is typical of holes.

The depth of penetration of electrons 1 into the junction naturally grows as they acquire more thermal energy.

3.2.2. Analysis of an equilibrium *pn* junction. The height of an equilibrium potential barrier depends on the difference between the electrostatic potentials in the p - and the n -layer (see Fig. 3.4b):

$$\Delta\varphi_0 = \varphi_{Ep} - \varphi_{En} \quad (3.1)$$

The potentials φ_{Ep} and φ_{En} are easy to obtain from Eqs. (2.11), substituting respectively $p = p_{p0}$ and $n = n_{n0}$ (the subscripts n and p identify the respective layers and the index 0 denotes the equilibrium state). Then,

$$\Delta\varphi_0 = \varphi_T \ln (n_{n0} p_{p0} / n_i^2) \quad (3.2a)$$

Setting $n_{n0} = N_d$ and $p_{p0} = N_a$, where N_d and N_a are the **effective** impurity concentrations, we get

$$\Delta\varphi_0 = \varphi_T \ln \frac{N_d N_a}{n_i^2} \quad (3.2b)$$

Other conditions being the same, the equilibrium height of the potential barrier will obviously increase with a decreasing intrinsic

concentration, or, which is the same, with an increasing width of the bandgap in the semiconductor. Substituting $N_d = 10^{19} \text{ cm}^{-3}$, $N_a = 10^{16} \text{ cm}^{-3}$ and the value of n_i for silicon (see Table 2.1) into Eq. (3.2b), we obtain the value of $\Delta\phi_0 = 33\phi_T \approx 0.83 \text{ V}$ at room temperature.

Using relation (2.9), replace in Eq. (3.2a) one of the **majority**-carrier concentrations, n_{n0} or p_{p0} , by the **minority**-carrier concentration p_{n0} or n_{p0} . We thus find out that the equilibrium height of the potential barrier depends on the relation between one-type carrier concentrations (electrons or holes) on both sides of the junction:

$$\Delta\phi_0 = \phi_T \ln (n_{n0}/n_{p0}) \quad (3.3a)$$

$$\Delta\phi_0 = \phi_T \ln (p_{p0}/p_{n0}) \quad (3.3b)$$

We shall use these variants of barrier representation later in analyzing the nonequilibrium state of the junction.

Expressions (3.2) show that the potential barrier height depends on temperature by way of the parameters ϕ_T and n_i . Resorting to (2.1) and (2.8), it is possible to obtain the expression for temperature sensitivity in the form

$$\varepsilon_0 = \frac{d(\Delta\phi_0)}{dT} = \frac{\Delta\phi_0 - \phi_g}{T} < 0 \quad (3.4)$$

For silicon, $\varepsilon_0 \approx -1.4 \text{ mV } ^\circ\text{C}^{-1}$ at $T = 300 \text{ K}$.

To calculate the equilibrium width of a junction, let us use the idealized distribution of charges (see Fig. 3.2c). With such a distribution (Fig. 3.6a), the charge densities on each side of the junction are constant (Fig. 3.6b): on the left side in the n^+ layer, $\lambda_n = qN_d$, and on the right in the p layer, $\lambda_p = -qN_a$. Substituting these expressions into Poisson's equation (2.45) and integrating it twice across the entire width of each of the two portions of the junction yields the linear distribution of field strength E and the quadratic distribution of potential ϕ (Fig. 3.6c and d).

The function $E(x)$ assumes the form

$$E(x) = \frac{qN_d}{\varepsilon_0\varepsilon} (l_n + x), \quad x \leq 0$$

$$E(x) = \frac{qN_a}{\varepsilon_0\varepsilon} (l_p - x), \quad x \geq 0$$

Equating the expressions for $E(x)$ at $x = 0$ (at the metallurgical boundary), we derive the relation between the width components of the junction in the n - and p -type layers:

$$l_n/l_p = N_a/N_d \quad (3.5)$$

In an asymmetric junction and especially in a one-sided junction of the n^+p type, an inequality $N_d \gg N_a$ is valid. Then $l_n \ll l_p$,

and, as mentioned earlier in the Subsec. 3.2.1, the total width of the junction is close in value to the width component of the high-resistance layer, $l_0 \approx l_p$.

The function $\varphi(x)$ has the form

$$\varphi(x) = \varphi_n - \frac{qN_d}{2\epsilon_0\epsilon} (x + l_n)^2, \quad x \leq 0$$

$$\varphi(x) = \varphi_p + \frac{qN_a}{2\epsilon_0\epsilon} (x - l_p)^2, \quad x \geq 0$$

Equating the expressions for $\varphi(x)$ at $x = 0$ and taking into account that $\varphi_n - \varphi_p = \Delta\varphi_0$, we find

$$\Delta\varphi_0 = \frac{qN_d}{2\epsilon_0\epsilon} l_n^2 + \frac{qN_a}{2\epsilon_0\epsilon} l_p^2$$

For asymmetric junctions, it is safe to omit one of the summands. Thus for an n^+p junction, where $l_n \ll l_p$, we can neglect the first

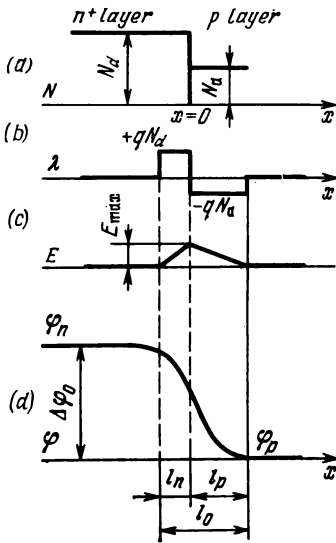


Fig. 3.6. Distribution of impurity concentrations (a), space charge density (b), field strength (c), and potential (d) in an abrupt n^+p junction

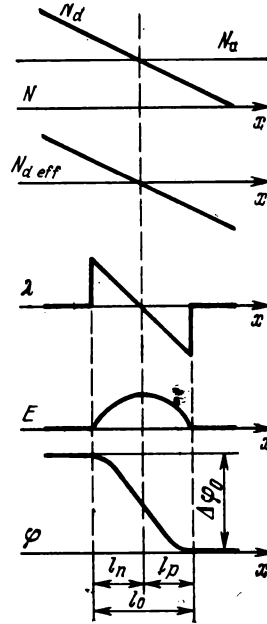


Fig. 3.7. Distribution of impurity concentrations, effective donor concentration, space charge density, field strength, and potential in a graded n^+p junction

term and put $l_p = l_0$. Omitting for uniformity the subscripts, we can now write the potential barrier width in an asymmetric junction

in the following form

$$l_0 = \sqrt{\frac{2\epsilon_0\epsilon\Delta\phi_0}{qN}} \quad (3.6)$$

where N is the impurity concentration in the **high-resistance** layer of the junction. Assuming $\Delta\phi_0 = 0.8$ V and $N = 10^{16}$ cm⁻³, we find that for silicon $l_0 \approx 0.3$ μ m.

Graded pn junctions are largely produced by high-temperature diffusion of impurities (see Sec. 6.5). The impurity distribution in this case closely follows the exponential law [see Eq. (2.76)]. The analysis of graded junctions, however, start from the assumption that the distribution of effective impurity concentrations over a short portion (within the junction width) obey the **linear law**. The space charge densities may then be regarded as linear functions of the coordinate x (Fig. 3.7). In this case the solution of Poisson's equation yields a quadratic function $E(x)$ and cubic function $\phi(x)$. Employing these functions in the same manner as above in analyzing the step junction, we can derive the equation for the width of an equilibrium graded junction:

$$l_0 = \sqrt[3]{\frac{9\epsilon_0\epsilon\Delta\phi_0}{qN'}} \quad (3.7)$$

where N' is the effective concentration gradient. Since the gradient is the same on either side of the junction, then the width l_0 is divided into equal lengths between the n - and p -layers. In this case the *junction is said to be symmetric*.

As for the height of the equilibrium potential barrier, we can find it with the aid of Eq. (3.2b), assuming that N_d and N_a are **effective** impurity concentrations at respective boundaries of the junction,

3.2.3. Analysis of a nonequilibrium pn junction. If an emf source V is applied across the p - and n -layers, the equilibrium state of the junction will be upset and an electric current will flow in the closed circuit. As mentioned above, the resistivity of the depletion layer is much higher than that of neutral layers. Therefore, *practically all the external voltage drops across the junction, which implies that a change in the potential barrier height is equal in value to the applied emf.*

With a positive voltage V applied to the p layer, the barrier height lowers (Fig. 3.8a):

$$\Delta\phi = \Delta\phi_0 - V \quad (3.8)$$

The voltage of such a polarity is called **forward**. At a negative potential on the p layer, the barrier height rises (Fig. 3.8b), for which reason the minus sign ahead of V in Eq. (3.8) should be changed for the plus. The voltage of this polarity is called **reverse**. In the further discussions, forward voltages will be regarded as positive, and reverse voltages as negative.

As the height of the potential barrier changes, its width and boundary carrier concentrations change too.

(Substituting $\Delta\varphi$ from Eq. (3.8) into Eq. (3.6), we find the unequilibrium barrier height

$$l = \sqrt{\frac{2\varepsilon_0\varepsilon(\Delta\varphi_0 - V)}{qN}} \quad (3.9)$$

The expression reveals that the *junction narrows at forward voltage* ($V > 0$) and *widens at reverse voltage* ($V < 0$). But at forward voltages in excess of 0.3 or 0.4 V (for silicon), formula (3.9) involves a substantial error because the idealization of the depletion layer as

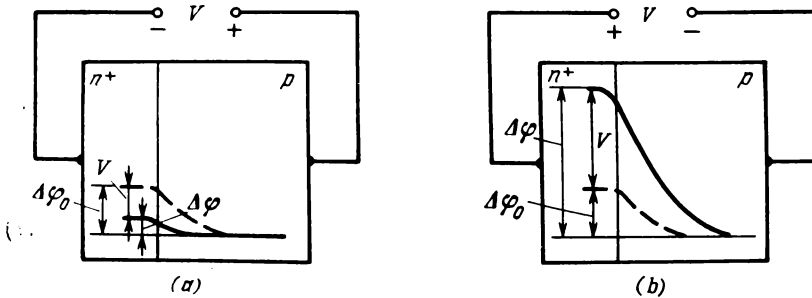


Fig. 3.8. Biasing of the junction to the forward (a) and the reverse (b) condition

adopted in Figs. 3.2c and 3.6 does not prove justifiable any longer. At reverse voltages, formula (3.9) remains quite acceptable. If the reverse voltage exceeds in magnitude the quantity $\Delta\varphi_0$ by a factor of 2 or 3 and more, a simplified variant of the formula may be used:

$$l \approx \sqrt{\frac{2\varepsilon_0\varepsilon}{qN} |V|} \quad (3.10)$$

For a graded junction, using Eq. (3.7) we obtain the expression in the general form

$$l = \sqrt[3]{\frac{9\varepsilon_0\varepsilon}{qN'}(\Delta\varphi_0 - V)} \quad (3.11a)$$

and at rather high reverse voltages,

$$l \approx \sqrt[3]{\frac{9\varepsilon_0\varepsilon}{qN'} |V|} \quad (3.11b)$$

It is obvious that the dependence $l(V)$ for the graded junction is weaker than for the abrupt junction.

A change in the height of a potential barrier involves, generally speaking, a change in all the four boundary concentrations. But since the concentrations of majority carriers substantially exceed the

minority-carrier concentrations, we have ground to assume that only the latter concentrations undergo changes. Based on this assumption, we replace the concentrations n_{p0} and p_{n0} by n_p and p_n in the right sides of Eqs. (3.3), and $\Delta\phi_0$ by $\Delta\phi$ in their left sides. Then, substituting $\Delta\phi_0$ from Eqs. (3.3), we can easily establish the relation between the boundary minority concentrations in the equilibrium and the nonequilibrium state:

$$n_p = n_{p0} e^{V/\phi_T} \quad (3.12a)$$

$$p_n = p_{n0} e^{V/\phi_T} \quad (3.12b)$$

At **forward voltages**, the boundary concentrations prove higher than at the equilibrium state. This means that excess carriers appear in each of the layers, that is, the process of *injection* takes place (see Fig. 2.26a). At **reverse voltages**, the boundary concentrations decrease against the equilibrium values, which is an indication that the *extraction* of carriers occurs (see Fig. 2.26b).

Find the **excess** concentrations on either side of the junction by subtracting respectively n_{p0} and p_{n0} from n_p and p_n :

$$\Delta n_p = n_{p0} (e^{V/\phi_T} - 1) \quad (3.13a)$$

$$\Delta p_n = p_{n0} (e^{V/\phi_T} - 1) \quad (3.13b)$$

Divide Eq. (3.13a) by (3.13b) and replace n_{p0} and p_{n0} by p_{p0} and n_{n0} using Eq. 2.9. Assuming that the latter concentrations are equal to respective impurity concentrations, we obtain

$$\Delta n_p / \Delta p_n = N_d / N_a \quad (3.14)$$

From the above expression it follows that in asymmetric junctions the excess carrier concentration in the high-resistance layer (with a low carrier concentration) is much higher than in the low-resistance layer. It can be said, therefore that in *asymmetric junctions the process of injection is unidirectional in character*: it is the carriers injected from the low-resistance (highly doped) layer into the high-resistance layer that play the main role.

The injecting layer (or a lower resistivity) is called an *emitter*, and the layer of a higher resistivity that receives carriers (minority carriers for this layer) injected from the emitter is called a *base*.

At reverse voltages (when extraction takes place), the boundary concentrations, according to (3.12), are lower than the equilibrium concentrations and can be as small as desired; the excess concentrations, according to (3.13), are negative and do not exceed in magnitude the equilibrium values n_{p0} and p_{n0} .

3.2.4. Current-voltage characteristic of a *pn* junction. Figure 3.9 shows a current flow model at the junction. In the general case

(Fig. 3.9a), the current comprises an *electron* and a *hole* component with subscripts *n* and *p* respectively, and each in turn includes an *injection* and a *recombination* component with respective superscripts “in” and “r”. Recombination components are due to recombination of carriers in the space charge region as they make way to an adjacent layer. Needless to say that the recombination components of

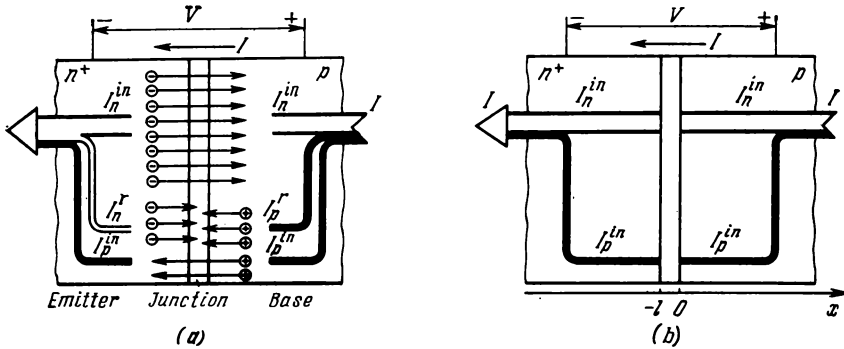


Fig. 3.9. Pattern of charge carrier transport in an n^+p junction, considering (a) and ignoring (b) recombination in the space charge region

electron and hole currents are equal; they sometimes play an important part (see Subsec. 4.4.3), but now in constructing an **ideal** current-voltage characteristic, we shall neglect these components and use the current flow pattern as illustrated in Fig. 3.9b.

The current flow model shows that electron and hole currents in both layers are equal, that is, they do not vary in the junction region. It can be said that, having omitted the recombination components, we set the junction width l equal to zero.

For the calculation of injection components, we shall avail ourselves of the fact that at the boundaries of the junction the electric field strength is equal to zero (see Fig. 3.6c) and, hence, the currents of injected carriers are *of the diffusion type* only [calculated by formulas (2.57)]. Write the boundary gradients of concentrations in the form of (2.69b), assuming that the junction width $l = 0$ (see above), in which case the base and emitter boundaries of the junction coincide. Then,

$$\left. \frac{d(\Delta n)}{dx} \right|_{x=0} = -\frac{\Delta n_p}{L_n}, \quad \left. \frac{d(\Delta p)}{dx} \right|_{x=0} = \frac{\Delta p_n}{L_p}$$

The plus sign for the hole gradient is due to injection of holes from the base into the emitter, that is, in the direction of the **negative** values of x (see Fig. 3.9). Substituting these gradients into (2.57) and using Eqs. (3.13) gives the electron and hole components of the

current in the form

$$j_n = -\frac{qD_n}{L_n} n_{p0} (e^{V/\varphi_T} - 1) \quad (3.15a)$$

$$j_p = -\frac{qD_p}{L_p} p_{n0} (e^{V/\varphi_T} - 1) \quad (3.15b)$$

The minus signs are due to the assumed direction of the x -axis from the negative to the positive pole of the emf V (see Fig. 3.9). From the physical viewpoint, currents flow from "plus" to "minus" and thus are positive. Summing up the quantities j_n and j_p , multiplying the result by the junction area S , and omitting the minus sign, we can write the expression for the I - V characteristic of a pn junction in the form

$$I = I_0 (e^{V/\varphi_T} - 1) \quad (3.16)$$

where

$$I_0 = \frac{qD_n S}{L_n} n_{p0} + \frac{qD_p S}{L_p} p_{n0} \quad (3.17)$$

Formula (3.16) is one of the most important in transistor electronics. The initial portion of the curve in Fig. 3.10 is plotted in relative units. The I - V characteristic definable by Eq. (3.16) is called *ideal* because it does not reflect many of the attendant factors. Real I - V characteristics differ from the ideal, but the corresponding expressions prove both less illustrative than (3.16) and too complex for practical use. In the analysis and calculation of semiconductor devices, therefore, at least at the first stage, it is usual to utilize (3.16) and then make the requisite corrections for errors.

The quantity I_0 that determines "the scale" of the I - V curve is termed *thermal current* because it owes its origin to heat and, as will be shown below, heavily depends on temperature. From Eq. (3.16) it is seen that at a sufficiently large reverse voltage, namely at $|V| > 3\varphi_T$, the value of reverse current is $-I_0$ and independent of voltage. It is thus safe to say that *thermal current determines the value of reverse current for an ideal I - V characteristic*¹. Besides,

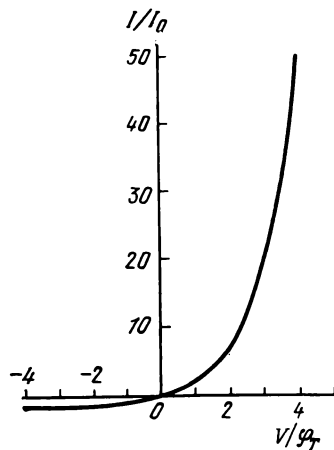


Fig. 3.10. I - V characteristic of an ideal pn junction (diode)

¹ For real I - V characteristics, as will be disclosed later in the text, thermal current is not always the main component of reverse current.

thermal current affects many parameters of pn junctions and transistors, for which reason we shall consider it in more detail.

In asymmetric junctions the current components j_n and j_p differ substantially because of the difference in excess concentrations [see Eq. (3.14)]. The components of thermal current differ accordingly. For an n^+p junction the main component is the **electron** current, that is, the augend on the right of Eq. (3.17). Write this component, replacing n_{p0} by n_i^2/p_{p0} according to Eq. (2.9) and setting $p_{p0} = N_a$. For generality, we omit subscripts. The expression reduces to the form

$$I_0 = (qDS/LN) n_i^2 \quad (3.18)$$

If $N = 10^{16} \text{ cm}^{-3}$ and $L = 10 \text{ } \mu\text{m}$, then the thermal current density for silicon at room temperature is $j_0 \approx 2 \times 10^{-10} \text{ A/cm}^2$. In modern integrated transistors base areas do not exceed $2 \times 10^{-5} \text{ cm}^2$, and emitter areas are as small as 10^{-6} cm^2 and below. At room temperature, therefore, the typical value of thermal current I_0 through a silicon pn junction in an IC may be taken to equal 10^{-15} A .

The temperature dependence of thermal current is determined by the quantity n_i^2 . Using Eq. (2.8), write the thermal current in the form

$$I_0 = I_{00} e^{-\Phi_g/\Phi_T} \quad (3.19)$$

where the quantity I_{00} may be regarded as being independent of temperature.

Differentiating (3.19) with respect to temperature and considering (2.1), it is easy to obtain the temperature coefficient (TC) of thermal current

$$\frac{dI_0/dT}{I_0} = \frac{\Phi_g/\Phi_T}{T} \quad (3.20)$$

At room temperature the TC for silicon is $16\% \text{ } ^\circ\text{C}^{-1}$. Since it depends on temperature, the TC proves inconvenient for use over a wide temperature range.

It is common to characterize the function $I_0(T)$ by the *temperature of current doubling*, T^* , that is, by the increment of temperature which causes the thermal current to increase twofold. From Eq. (3.19), we can obtain an approximate expression:

$$T^* \approx \left(\frac{\Phi_{T0}}{\Phi_g} \ln 2 \right) T_0$$

where T_0 is the average temperature for a certain temperature range. At room temperature, T^* for silicon is equal to about 5°C . If the thermal current at T_0 is known, then its value at any other temperature T may approximately be estimated by the relation

$$I_0(T) = I_0(T_0) 2^{\Delta T/T^*} \quad (3.21)$$

where $\Delta T = T - T_0$. It can be readily ascertained that comparatively small changes in temperature cause the thermal current to vary by a few orders of magnitude. Thus at 125°C, the thermal current through an IC silicon junction may reach 10^{-8} A and over.

3.2.5. Forward current-voltage characteristic. At voltages $V > 0$, the I - V curve is so steep that it is difficult to obtain the desired current at a specified voltage: the slightest variation in voltage tends to change the current to a substantial degree. That is why it is typical for *pn junctions to operate at the specified forward current*. To examine the V - I relation, write Eq. (3.16) for the I - V characteristic in the form

$$V = \varphi_T \ln \left(\frac{I}{I_0} + 1 \right)$$

For silicon junctions showing a negligible thermal current, the following expression may be applicable

$$V = \varphi_T \ln (I/I_0) \quad (3.22)$$

If forward currents vary by a factor of 10^2 and over, the forward voltage may vary heavily too. But in practice the current range can rarely be so wide. Over the common operating range the forward voltage alters but very insignificantly.

For example, if $I_0 = 10^{-15}$ A and forward currents lie in the range 10^{-3} to 10^{-4} A (known as the *normal current range*), the voltage V varies merely from 0.69 to 0.64 V. At other currents I_0 spanning the range 10^{-5} to 10^{-6} A (known as the *microampere region*, or μ A-range), the voltage changes from 0.57 to 0.52 V.

Thus, forward voltages differ somewhat depending on the range of currents, but for a definite current range these voltages may be thought to be constant and regarded as a kind of parameter of the forward-biased silicon junction. Let us designate this parameter as V^* and call it the *voltage of a forward-biased junction*. At room temperature, we shall set this voltage at 0.7 V over the normal current range and at 0.5 V in the microampere range.

If the forward voltage is merely 0.1 V (or $4 \varphi_T$) below the voltage V^* , the junction may practically be considered still nonconductive since the currents at such a voltage are a few orders of magnitude smaller than the rated values. The parameter $V^* - 0.1$ V may conditionally be called the "cut-off" voltage of the junction. This bias voltage is well illustrated in Fig. 3.11a. At lower voltages, down to zero, the I - V curve merges into the x -axis to form what is called a "heel" of the I - V characteristic.

The voltage V^* depends on temperature at an invariable current. This temperature dependence is due to the components φ_T and I_0 entering into Eq. (3.22). Differentiating (3.22) with respect to tem-

perature and considering (3.20), it is easy to obtain the expression for the temperature sensitivity of forward voltage:

$$\varepsilon^* = dV^*/dT = -(\varphi_g - V^*)/T \quad (3.23)$$

The parameter ε^* is weakly dependent on temperature because with an increase in T the quantity V^* in the numerator decreases accordingly. For silicon, ε^* lies between $-1.5 \text{ mV } ^\circ\text{C}^{-1}$ in the normal current range and $-2 \text{ mV } ^\circ\text{C}^{-1}$ in the μA -range.

As seen, with rising temperature the forward voltage on the pn junction drops off (dash lines in Fig. 3.11a). At $T = 150^\circ\text{C}$, the forward voltage may be 0.2 to 0.25 V below the rating.

An important feature of the ideal I - V characteristic is the inverse relationship between the forward voltage and thermal current: *the lower the thermal current, the higher the forward voltage, and vice versa*.

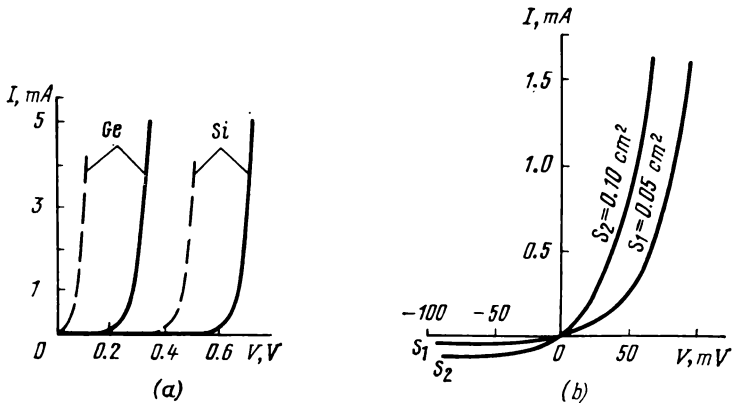


Fig. 3.11. I - V characteristics of ideal diodes (pn junctions) with various band-gap widths (a) and various junction areas (b)

versa. For this reason, in germanium junctions, where the current I_0 is 10^6 times as large as that in silicon junctions, the forward voltages are lower by 0.35 V, other conditions being the same; they commonly vary from 0.25 to 0.15 V, as shown in Fig. 3.11a, depending on the working current range. Thus, the "cut-off" voltage is close to zero and the parameter V^* has no meaning, so there is more sense in assuming $V^* \approx 0$.

One more consequence of the inverse relation between V and I_0 is that the forward voltage decreases with an increase in the junction area (Fig. 3.11b).

A feature typical of the *real* I - V characteristic is an ohmic drop of voltage across the base layer. Indeed, if the base layer has a rather high resistance r_b , the external voltage, generally speaking, *does not*

drop completely across the pn junction, but spreads out between the junction and base layer. The forward voltage given by Eq. (3.22) will then be expressed as the sum of terms:

$$V = \varphi_T \ln (I/I_0) + Ir_b \quad (3.24)$$

Since the addend (ohmic drop) is a linear function of current, and the augend is a logarithmic function, it is clear that at a fairly large

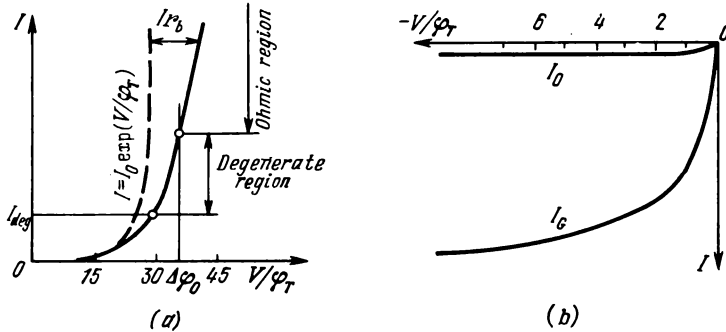


Fig. 3.12. Forward (a) and reverse (b) current-voltage characteristic of a real silicon diode (pn junction)

current the exponential I - V characteristic always degenerates, that is, becomes flatter (see Fig. 3.12a). Such degeneracy sets in at a current

$$I_{deg} = \varphi_T/r_b$$

The base resistance in a small-area junction may range into the tens of ohms, so the I - V characteristic may tend to degenerate at comparatively small currents, from 0.2 to 0.5 mA.

If the forward voltage exceeds $\Delta\varphi_0$, then the potential barrier height practically drops to zero and the I - V characteristic becomes quasilinear:

$$V = \Delta\varphi_0 + Ir_b$$

This portion of the I - V curve is called *ohmic*. The strict linearity does not exist because of *base resistance modulation*—an increase in the base conductivity due to high concentrations of excess carriers at large currents. Under these conditions, the relation $I \sim V^2$ becomes valid for the I - V curve.

One of the important parameters of the forward current-voltage characteristic is an *incremental resistance across the junction*. For the initial (nondegenerate) portion this resistance can be readily obtained from Eq. (3.22):

$$r_{pn} = dV/dI = \varphi_T/I \quad (3.25)$$

The physical meaning of this parameter becomes clear if we replace the differentials dV and dI by finite increments. Then

$$\Delta V = \Delta I r_{pn}$$

Hence, r_{pn} is the resistance to the current increments, ΔI , which are small as against the dc component I that determines the value of r_{pn} .

The typical value of r_{pn} , which it is advisable to remember, is 25 Ω at I equal to 1 mA. The recalculation of resistance at other values of current does not present difficulties. It should be pointed out that the junction resistance rises sharply over the μA -range. Thus at $I = 5 \mu\text{A}$, $r_{pn} = 5 \text{ k}\Omega$.

3.2.6. Reverse current-voltage characteristic. As noted earlier, the real reverse current through a silicon junction by far exceeds the value of I_0 as predicted by Eq. (3.17). The cause of this is first of all the generation of electron-hole pairs in the space charge region of the reverse-biased junction. The component of reverse current that results from this phenomenon is called a *thermally generated current*.

The processes of generation and recombination of carriers occur in all the portions of a diode, both in neutral n - and p -layers and in the junction region. In the equilibrium state, the rates of generation and recombination are equal throughout, and so directional flows of carriers are absent. The reverse voltage applied to the junction causes depletion of carriers in the junction region. Recombination here slows down and the process of generation proves nonequilibrium. The excess carriers being generated move under the action of the electric field into the n layer and holes into the p layer. These flows of carriers are responsible for the thermally generated current I_G .

To determine the current I_G , one must know the carrier generation rate given by (2.36). For simplicity, trap levels are assumed to lie exactly in the middle of the bandgap (in which case $n_t = p_t = n_i$) and the free carrier concentrations in the junction region are taken to equal zero at the reverse voltage ($n = p = 0$). Given such conditions, Eq. (2.36) yields the generation rate in the form

$$dn/dt = n_i/\tau$$

where $\tau = \tau_p + \tau_n$ is the total lifetime. Multiplying the generation rate by the pn junction volume Sl and the elementary charge q on each particle, we get

$$I_G = (qSl/\tau) n_i \quad (3.26)$$

For a silicon junction with $\tau = 0.1 \mu\text{s}$, $l = 0.5 \mu\text{m}$, and $S = 5 \times 10^{-6} \text{ cm}^2$, we find that $I_G \approx 10^{-11} \text{ A}$. This value is four orders of magnitude higher than the value of thermal current calculated earlier. In semiconductors showing high intrinsic densi-

ties n_i , the difference between I_G and I_0 is smaller since $I_0 \sim n_i^2$, while $I_G \sim n_i$. Thus in a germanium junction, both currents are of the same order of magnitude.

Apart from its high value, the thermally generated current differs from the thermal current by its *dependence on reverse voltage* (Fig. 3.12b). This dependence follows from Eqs. (3.9) and (3.10) which determine the junction width.

As regards the temperature dependence of current I_G , it can be easily estimated by substituting n_i from (2.8) into (3.26). The thermally generated current is then described by

$$I_G = I_{G0} e^{-\varphi_g/2\varphi_T} \quad (3.27)$$

where I_{G0} is the quantity weakly dependent on temperature. Comparing Eq. (3.27) with Eq. (3.19), we can see that I_G is less dependent on T than I_0 . In particular, the TC/I_G will be half as large (about $8\% \text{ } ^\circ\text{C}^{-1}$) and the temperature of I_G doubling will be twice as high (about 10°C).

3.2.7. Junction breakdown. There are basically three kinds (or mechanisms) of breakdown in *pn* junctions at fairly large reverse voltages: *tunnel*, *avalanche*, and *thermal*. The first two stem from an

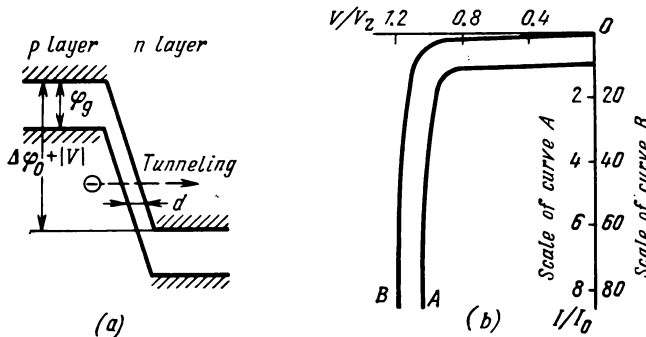


Fig. 3.13. Tunnel breakdown

(a) band diagram; (b) reverse I - V characteristic of diode under conditions of tunnel breakdown

increase in the strength of an electric field in the junction, and the third from an increase in the dissipated power and thus in temperature.

What underlies the mechanism of tunnel breakdown is the tunnel effect responsible for the penetration of electrons through a thin potential barrier (see p. 59). In the given case the barrier height is meant to be the bandgap width φ_g , and its thickness is understood to be the distance d between the opposite bands (Fig. 3.13a). The

analysis yields the following approximate expression for breakdown voltage:

$$V_Z = 1/2 \epsilon_0 \epsilon \mu_b E_{br}^2 \rho_b \quad (3.28a)$$

where E_{br} is the breakdown strength of the field (for silicon, $E_{br} = 4 \times 10^5$ V/cm); the subscript b relates to the base layer. In practice, semiempirical formulas are used. Thus for silicon,

$$V_Z = 40\rho_n + 8\rho_p \quad (3.28b)$$

where ρ_n and ρ_p are the resistivities of respective layers in Ω cm.

Both formulas show that the tunnel breakdown voltage is proportional to the base resistivity. For this reason, *to increase the breakdown voltage it is necessary to use a sufficiently high-resistance material for the base*. The general form of the reverse characteristic in tunnel breakdown is displayed in Fig. 3.13b.

The mechanism of avalanche breakdown depends on the multiplication of carriers in a strong electric field that acts in the junction

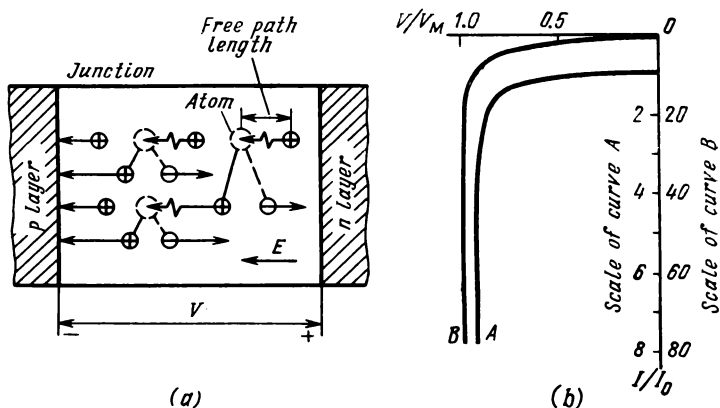


Fig. 3.14. Avalanche breakdown

(a) electron multiplication model; (b) reverse I - V characteristic of diode under conditions of avalanche breakdown

region¹. An electron or hole accelerated by the field acquired enough energy in its free path to be able to rupture one of the covalent bonds of a neutral atom in the semiconductor, thereby generating an electron-hole pair by impact ionization. These carriers excite new electron-hole pairs, and so on (Fig. 3.14a). The reverse current here naturally grows. At a sufficiently high strength of the field, in which the initial pair of carriers excites on the average more than one pair of

¹ The field strength in the junction can roughly be estimated as a quotient of the reverse voltage by the junction width. Setting $|V| = 5$ V and $l = 1$ μ m, we get $E = 50$ kV/cm.

new carriers, ionization assumes an *avalanche* form, analogous to the avalanche breakdown in a gas discharge. In this process only the external resistance sets limits on the current.

The I - V characteristic accounting for carrier multiplication until breakdown is described by a semiempirical formula

$$M = \frac{I}{I_0} = \frac{1}{1 - (V/V_M)^n} \quad (3.29)$$

where M is the *coefficient of impact ionization*, I and V are the absolute values of reverse current and voltage respectively, and V_M is the avalanche breakdown voltage (at which $M = \infty$). The values of exponent n for silicon appear in Table 3.1. The general form of the reverse characteristic in avalanche breakdown is shown in Fig. 3.14*b*.

The avalanche breakdown voltage is related to the base resistivity by a semiempirical formula

$$V_M = a\rho_b^m \quad (3.30)$$

where ρ_b has a dimension of Ω cm; the values of a and m are given in Table 3.1. The dependence $V_M(\rho_b)$ is weaker than $V_Z(\rho_b)$, there-

Table 3.1

Parameters of Avalanche Breakdown

Material	Conductivity type of base	n	m	a
Silicon	Electron	5	86	0.65
	Hole	3	23	0.75
Germanium	Electron	3	83	0.60
	Hole	5	52	0.60

fore at high values of ρ_b (when $V_M < V_Z$) the mechanism of breakdown is of the avalanche nature, while at low values of ρ_b (when $V_Z < V_M$) the tunnel mechanism is operative. The limiting value of breakdown voltage is equal to about 5 V. Above this value breakdown is of the avalanche type, and below this value, it is of the tunnel type.

A distinguishing feature of both types of breakdown is that the *TC* of avalanche breakdown voltage is opposite in sign to the *TC* of tunnel breakdown voltage (Fig. 3.15). The reason is that the tunnel breakdown voltage is directly dependent on the bandgap width, and

therefore a decrease in ϕ_g with rising temperature [see Eq. (2.4)] causes a decrease in V_Z . The avalanche breakdown voltage is inversely dependent on carrier mobility, and so a drop in μ with growing temperature [see Eqs. (2.20)] leads to an increase in V_M .

Both types of breakdown find practical uses in *silicon reference diodes* intended to stabilize voltage; in the breakdown region, as

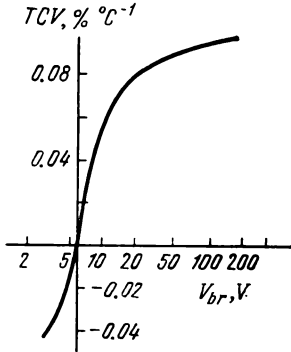


Fig. 3.15. TCV versus breakdown voltage

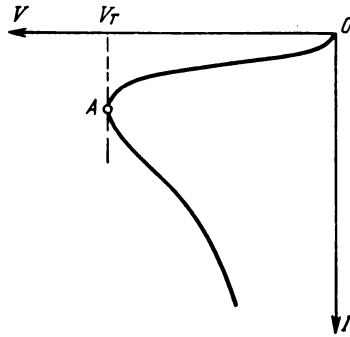


Fig. 3.16. Reverse I - V characteristic of a diode under conditions of thermal breakdown

seen from Figs. 3.13 and 3.14, the voltage weakly depends on current changes, hence the diode can function as a voltage stabilizer (see Sec. 9.10).

The voltage is not strictly constant because the I - V curve in the breakdown portion is not entirely vertical. The slope of the curve is described by an incremental resistance $r_{st} = dV/dI$.

For tunnel breakdown,

$$r_{st} \approx \frac{V_Z}{I} E_{br} 10^{-7} \quad (3.31a)$$

For avalanche breakdown,

$$r_{st} \approx \frac{V_M}{I} \left(1 - \frac{V}{V_M} \right) \quad (3.31b)$$

For example, if $V_Z = 3$ V, $I = 2$ mA, $E_{br} = 4 \times 10^5$ V/cm, $V_M = 10$ V, and $V/V_M = 0.98$, then we get $r_{st} \approx 60 \Omega$ and $r_{st} \approx 100 \Omega$ for the tunnelling-effect and avalanche-effect devices respectively.

With an increase in current, r_{st} decreases. Depending on the junction area, the minimum values of r_{st} lie in the range from 2-10 Ω (for large areas) to 20-50 Ω (for small areas).

Thermal breakdown results from self-heating of the junction as the reverse current flows through it. As the temperature rises, reverse currents grow sharply [see Eqs. (3.19) and (3.25)] and so does

the energy dissipated in the junction, producing a further increase in temperature, and so on. A distinctive feature of the I - V characteristic in thermal breakdown is a portion of *negative* incremental resistance, $dV/dI < 0$ (below point A in Fig. 3.16). The equation for thermal breakdown voltage (at A) has the following structure:

$$V_T \approx 3/(\varphi_g R_t I_{rev}) \quad (3.32)$$

where R_t is the *thermal resistance at the junction*¹; and I_{rev} is the reverse current at room temperature. Substituting the typical values of $R_t = 0.5^\circ \text{C/mW}$ and $I_{rev} = 10^{-10}$ A, we obtain $V_T \approx 6 \times 10^7$ V for silicon. This voltage is much higher than the voltage of avalanche and tunnel breakdown.

From the above it can be inferred that thermal breakdown is not an independent phenomenon: it can begin to build up only *when the reverse current has grown high enough as a result of avalanche or tunnel breakdown* (for example, at $I_{rev} = 1$ mA, the voltage V_T drops to 6 V).

3.3. Transient Behavior of *pn* Junctions

A semiconductor diode is inert to rather fast variations in current or voltage because a new distribution of carriers does not set in at an instant. As known, the externally applied voltage changes the junction width (see Subsec. 3.2.3), and thus it changes the value of space charges in the junction. Besides, the same voltage causes injection or extraction of carriers and thus changes the charge in the base region (the role of charges in the emitter is insignificant). Consequently, a diode has a capacitance which may be thought of as the one being in parallel with the *pn* junction.

It is customary to divide this capacitance into a *barrier capacitance* (junction or depletion-layer capacitance), which reflects the redistribution of charges in the junction, and a *diffusion capacitance*, which reflects the redistribution of charges in the base. Such a division is in general conditional but convenient from the practical viewpoint, since the proportion of both capacitances is different with different polarities of applied voltage. In the forward-biased junction, it is the excess carriers in the base that play the major part (that is, the diffusion capacitance). At the reverse voltage, the amount of excess charges in the base is small and so the barrier ca-

¹ The thermal resistance is a proportionality coefficient in the relation $T_{pn} - T_{am} = R_t P$, where T_{pn} is the junction temperature, T_{am} is the ambient temperature, and P is the power scattered at the junction. This coefficient depends on the thermal conductivity and geometry of a crystal and is commonly estimated experimentally.

capacitance plays the main role. Consider first both of these capacitances and the transient processes proper.

3.3.1. Barrier capacitance. Assume the charge distribution over the junction to be such as shown in Fig. 3.6, that is, of the abrupt type. As before, we shall regard the junction as being of an asymmetric, n^+p type. Based on these assumptions, we may consider the width of a negative space charge region in the p base equal to the entire width of the junction, $l_p = l$ [see description of Eq. (3.5)]. Write the modulus of this charge:

$$|Q| = qNSl$$

where N is the impurity concentration in the base, and S is the junction area. The same (but positive) charge will be in the emitter n^+ layer.

Assume these charges are located on the plates of an imaginary capacitor. The capacitance of a capacitor is commonly expressed as Q/V . In the given case, however, the charge is attributed not only to the external voltage V but also to the equilibrium barrier height $\Delta\phi_0$ [see Eq. (3.9)]. So the charge $|Q|$ need be divided by the quantity $\Delta\phi = \Delta\phi_0 - V$. Transistor electronics more often uses a **differential** barrier capacitance obtained by differentiating $|Q|$ with respect to the voltage $\Delta\phi$. Considering Eq. (3.9), we derive the expression for barrier capacitance per unit area:

$$C_{b0} = \sqrt{\frac{0.5\epsilon_0 eqN}{\Delta\phi_0 - V}} \quad (3.33)$$

At forward voltages ($V > 0$), the barrier capacitance rises, but this increase in capacitance is not entirely congruous to the derived expression, the causes of incongruity being the same for Eq. (3.9)¹. Therefore, Eq. (3.33) is practically suitable only for reverse voltages. In this case, it is more convenient to use the modulus of reverse voltage $|V|$. Then,

$$C_{b0} = \sqrt{\frac{0.5\epsilon_0 eqN}{\Delta\phi_0 + |V|}} \quad (3.34a)$$

If the reverse voltage exceeds the value of $\Delta\phi_0$ by a factor of 2 and above, a simplified expression may be quite acceptable:

$$C_{b0} = \sqrt{\frac{qe_0eN}{2|V|}} \quad (3.34b)$$

For the silicon ($\epsilon = 12$),

$$C_{b0} \approx 3 \times 10^{-16} \sqrt{\frac{N}{|V|}}$$

¹ For the rated forward voltage V^* , it is usual to set $C_b = 1.5$ to $2 C_b(0)$, where $C_b(0)$ is the capacitance at the zero voltage across the junction.

For example, if $N = 10^{16} \text{ cm}^{-3}$ and $|V| = 4 \text{ V}$, the typical value of C_{b0} is approximately equal to $1.5 \times 10^{-8} \text{ F/cm}^2$ (150 pF/mm^2). With the junction area $S = 10^{-2} \text{ mm}^2$, the barrier capacitance will amount to 1.5 pF .

The dependence of barrier capacitance on voltage is illustrated in Fig. 3.17 for an abrupt and a graded *pn* junction. The capacitance

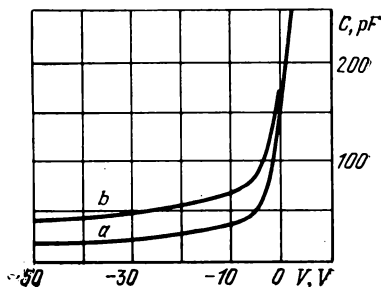


Fig. 3.17. Barrier capacitance of an abrupt (a) and a graded (b) junction as a function of reverse voltage

for the latter junction is found in the same fashion, but with the use of Eq. (3.11a) instead of Eq. (3.9). The graded junction shows a weaker dependence of its capacitance on voltage than the step junction.

3.3.2. Diffusion capacitance. The diffusion capacitance is due to changes of charges in the quasineutral base layer. Indeed, with the injection of minority carriers into the base of a *p⁺n* junction, the base becomes richer in both electrons and compensating holes. An increment in their charges divided by the voltage increment at the junction is just the diffusion capacitance.

Since the excess charges of electrons and holes are equal, let us find one of them, namely, the charge of electrons. We shall proceed from distribution equation (4.2) which, in distinction to Eq. (2.68), is valid for the finite base thickness [in deriving Eq. (2.68), the assumption was that $w = \infty$, i.e. $w \gg L$]. The excess charge on electrons will thus be written in the form

$$\Delta Q = qS \int_0^w \Delta n_b(x) dx = I\tau \left(1 - \operatorname{sech} \frac{w}{L}\right) \quad (3.35)$$

using Eq. (2.66) after integration.

Differentiating ΔQ with respect to V and considering Eq. (3.25) gives the expression for diffusion capacitance in the general form

$$C_{dif} = \frac{I\tau}{qT} \left(1 - \operatorname{sech} \frac{w}{L}\right) \quad (3.36)$$

For a thick base, where $w \gg L$ and $\text{sech}(w/L) \approx 0$, we obtain

$$C_{dif} = \frac{I\tau}{\Phi_T} \quad (3.37a)$$

Thus if $\tau = 1 \mu\text{s}$ and $I = 1 \text{ mA}$, then $C_{dif} = 0.04 \mu\text{F}$.

For a thin base, where $w < L$ and $\text{sech}(w/L) \approx 1 - 0.5 (w/L)^2$, the diffusion capacitance, considering Eq. (2.66), may be written thus

$$C_{dif} = \frac{I}{\Phi_T} \frac{w^2}{2D} = \frac{It_D}{\Phi_T} \quad (3.37b)$$

where

$$t_D = \frac{w^2}{2D} \quad (3.38)$$

is the mean *diffusion time*, that is, the mean transit time the carriers take to diffuse through a thin base. If $w = 5 \mu\text{m}$ and $D = 36 \text{ cm}^2/\text{s}$, then $t_D \approx 3.5 \text{ ns}$. At the same value of current as that taken in the preceding example (1 mA), $C_{dif} \approx 140 \text{ pF}$, or one three-hundredth that for the thick base.

The diffusion time proves to be as fundamental a parameter of semiconductor devices as the lifetime. It is easy to see that both equations (3.37) are the same in structure; they only differ in that the parameter t_D for the thin base replaces the parameter τ for the thick base.

Comparing the diffusion capacitance (DC) with the barrier capacitance (BC) we can make the following conclusions:

- (a) unlike BC, DC is independent of junction area;
- (b) DC is a function of current, whereas BC is a function of voltage;
- (c) at I of the order of 1 mA and higher, DC is a few orders of magnitude larger than BC, so it is quite safe to ignore the latter quantity;
- (d) in the microampere range covering the currents in the order of 1 μA and below, DC becomes comparable to BC.
- (e) with the reverse bias maintained on a silicon junction, when the current does not exceed 10^{-3} – $10^{-4} \mu\text{A}$, DC is close to zero and can be neglected with good reason.

3.3.3. General characteristic of transients. For rather small increments in voltage (less than Φ_T), the pn junction may be regarded as an incremental resistance [see Eq. (3.25)] shunted by diffusion and barrier capacitances, one of which may usually be neglected (see the conclusions at the end of Subsec. 3.3.2). The transient in this case is the same as that in the common parallel RC circuit. This process is of no special interest since diodes generally operate at voltages higher than Φ_T , in which case nonlinear properties of junctions begin to show up.

Large variations in current and voltage produce changes in the incremental resistance, barrier capacitance and, primarily, in the diffusion capacitance. Therefore, where large signals are present, the analysis involving the diffusion capacitance and incremental resistance turns out to be unsuitable; the analysis of the nonlinear *RC*

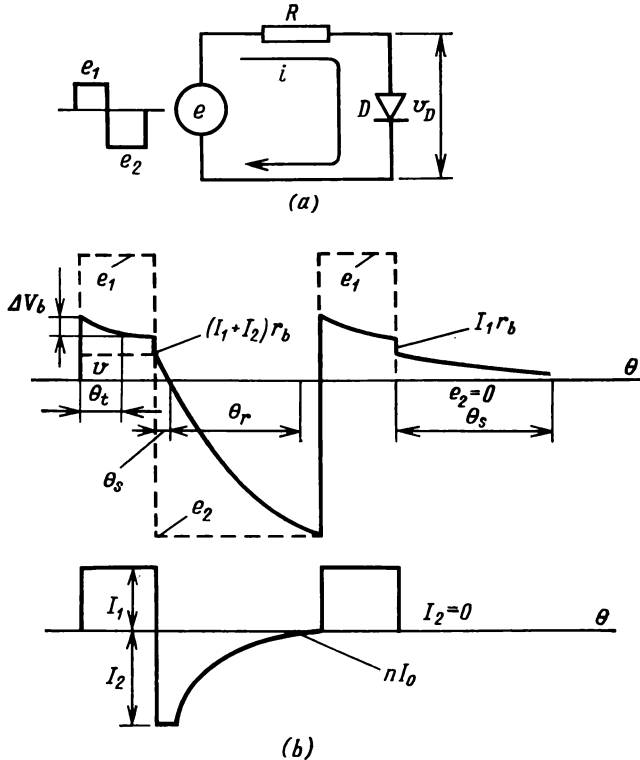


Fig. 3.18. Transient processes in a diode
(a) switching circuit; (b) time diagrams for switching and turn-off transients

circuit proves far from being a simpler approach than that relying directly on continuity equations.

The limiting case of a large signal is **switching** of the junction from the reverse to the forward state, and vice versa. It is exactly this case that is considered below. For simplicity, we shall take no account of barrier capacitance which delays somewhat the transient processes but does not change their nature.

The analysis of transients is commonly made for a step input (Fig. 3.18), with the *pn* junction alternately operated in the forward and reverse directions (in a particular case, the reverse switching

may be absent, that is, the diode may simply be cut off, $e_2 = 0$).

In switching the junction from the reverse to the forward state and vice versa, its transient characteristic displays the following regions (Fig. 3.18b):

- (1) transition interval θ_t , *setup of forward voltage* at a given forward current;
- (2) storage interval θ_s , *removal of excess carriers* in the base at a given reverse current;
- (3) recovery interval θ_r , *recovery of reverse current* at a given reverse voltage.

Consider these regions in sequence.

3.3.4. Setup of forward voltage. In this region, the specified value is considered to be the forward current:

$$I_1 = \frac{e_1 - V_d}{R} \approx \frac{e_1 - V^*}{R} = \text{constant}$$

The voltage V_d comprises two components, the voltage V at the pn junction proper (in the space charge region) and voltage V_b across the quasineutral base layer.

The transient characteristic for the first component v is determined from transient diffusion equation (2.64) with the use of Eq. (3.13a). For a thick base,

$$v(\theta) = \varphi_T \ln \left(\frac{I_1}{I_0} \operatorname{erf} \sqrt{\theta} + 1 \right) \quad (3.39)$$

where $\theta = t/\tau$ is the relative time, and $\operatorname{erf} \sqrt{\theta}$ is the error function, (see p. 72). The approximation of $\operatorname{erf}(x)$ relies on the similarity of this function to the exponential function $1 - e^{-x}$. Therefore, $\operatorname{erf} \sqrt{\theta} \approx \sqrt{1 - e^{-\theta}}$. At $\theta < 0.5$, $\operatorname{erf} \sqrt{\theta} \approx \sqrt{\theta}$, and at $\theta > 0.5$, $\operatorname{erf} \sqrt{\theta} \approx 1 - (1/2)e^{-\theta}$. Disregarding the unity in Eq. (3.39), we use the approximation $\operatorname{erf} \sqrt{\theta} \approx \sqrt{\theta}$ and equate the right side of Eq. (3.39) to steady-state equation (3.22) multiplied by 0.9 to obtain the transition time at 90% of its final value:

$$\theta_{t1} \approx \left(\frac{I_1}{I_0} \right)^{-0.2} \quad (3.40)$$

Thus if $I_1/I_0 = 10^{10}$, then $\theta_{t1} \approx 0.01$.

The transient characteristic for the second component v_b depends on the modulation of base resistance (see p. 91). At the start of current flow, I_1 , the base has only equilibrium charges. Further, as the base accumulates additional, excess charges due to injection, its resistance r_b diminishes, which entails a decrease in the voltage drop $I_1 r_b$. The charge accumulation and, hence, the decrease in voltage v_b depend on the mean lifetime τ .

A simplified analysis gives the following relation for a thick base:

$$v_b(\theta) = V_b(0) \left\{ 1 - \frac{L}{w} \ln \left[\frac{1}{2} \frac{I_1}{I_0} \left(\frac{\rho_{b0}}{\rho_i} \right)^2 + 1 \right] \operatorname{erf} \sqrt{\theta} \right\} \quad (3.41)$$

where $V_b(0) = I_1 r_{b0}$ is the initial voltage across the **nonmodulated** base resistance r_{b0} . We define the initial surge ΔV_b (see Fig. 3.18b) as the difference between $V_b(0)$ and $V_b(\infty)$:

$$\Delta V_b = V_b(0) \frac{L}{w} \ln \left[\frac{1}{2} \frac{I_1}{I_0} \left(\frac{\rho_{b0}}{\rho_i} \right)^2 + 1 \right] \quad (3.42)$$

For example, if $I_1/I_0 = 10^{10}$, $\rho_{b0}/\rho_i = 10^{-5}$ (at $\rho_{b0} = 2 \Omega \text{ cm}$), and $w/L = 2$, then $\Delta V_b \approx 0.2 V_b(0)$.

The *transition time* (at a level $V_b(\infty) + 0.1 \Delta V_b$) is found to be equal to

$$\theta_{t2} \approx 1.6 \quad (3.43)$$

As seen, $\theta_{t2} \gg \theta_{t1}$. Hence, it is safe to neglect the transient $v(\theta)$, considering that the stationary voltage, as given by Eq. (3.22), sets in **immediately** after arrival of the pulse I_1 .

We have assumed above that the length of pulse I_1 exceeds θ_{t2} . Where narrower pulses are involved, charges have no time to build up completely and the relative value of voltage surge diminishes; it also does so when the forward current I_1 decreases, as is apparent from Eq. (3.42).

3.3.5. Removal of excess carriers. After switching the diode from the forward to the reverse state, the charge stored up in the base cannot change in an instant. This is particularly obvious if we regard the diode as being shunted by the diffusion capacitance indicative of the presence of the above charge. Like any other capacitance, this capacitance recharges gradually under the action of the current flow. For the same reason the voltage across the junction does not change in an instant after switching; it then drops smoothly to zero and further grows, with its sign reversed, to the steady-state value e_2 (see Fig. 3.18b). The *storage time* for the charge in the base is the time interval between the moment of reverse switching and the moment at which the forward voltage v across the junction drops to zero¹.

If at the stage of charge removal the relation $v \ll e_2$ holds true, then the reverse current proves equal to the specified value:

$$-I_2 = \frac{-e_2 + v}{R} \approx \frac{-e_2}{R} = \text{constant}$$

¹ Analysis shows that at $v = 0$ a certain residual charge still remains in the base, but its removal takes place in the next stage of the transient process.

may be absent, that is, the diode may simply be cut off, $e_2 = 0$).

In switching the junction from the reverse to the forward state and vice versa, its transient characteristic displays the following regions (Fig. 3.18b):

(1) transition interval θ_t , *setup of forward voltage* at a given forward current;

(2) storage interval θ_s , *removal of excess carriers* in the base at a given reverse current;

(3) recovery interval θ_r , *recovery of reverse current* at a given reverse voltage.

Consider these regions in sequence.

3.3.4. Setup of forward voltage. In this region, the specified value is considered to be the forward current:

$$I_1 = \frac{e_1 - V_d}{R} \approx \frac{e_1 - V^*}{R} = \text{constant}$$

The voltage V_d comprises two components, the voltage V at the pn junction proper (in the space charge region) and voltage V_b across the quasineutral base layer.

The transient characteristic for the first component v is determined from transient diffusion equation (2.64) with the use of Eq. (3.13a). For a thick base,

$$v(\theta) = \varphi_T \ln \left(\frac{I_1}{I_0} \operatorname{erf} \sqrt{\theta} + 1 \right) \quad (3.39)$$

where $\theta = t/\tau$ is the relative time, and $\operatorname{erf} \sqrt{\theta}$ is the error function, (see p. 72). The approximation of $\operatorname{erf}(x)$ relies on the similarity of this function to the exponential function $1 - e^{-x}$. Therefore, $\operatorname{erf} \sqrt{\theta} \approx \sqrt{1 - e^{-\theta}}$. At $\theta < 0.5$, $\operatorname{erf} \sqrt{\theta} \approx \sqrt{\theta}$, and at $\theta > 0.5$, $\operatorname{erf} \sqrt{\theta} \approx 1 - (1/2)e^{-\theta}$. Disregarding the unity in Eq. (3.39), we use the approximation $\operatorname{erf} \sqrt{\theta} \approx \sqrt{\theta}$ and equate the right side of Eq. (3.39) to steady-state equation (3.22) multiplied by 0.9 to obtain the transition time at 90% of its final value:

$$\theta_{t1} \approx \left(\frac{I_1}{I_0} \right)^{-0.2} \quad (3.40)$$

Thus if $I_1/I_0 = 10^{10}$, then $\theta_{t1} \approx 0.01$.

The transient characteristic for the second component v_b depends on the modulation of base resistance (see p. 91). At the start of current flow, I_1 , the base has only equilibrium charges. Further, as the base accumulates additional, excess charges due to injection, its resistance r_b diminishes, which entails a decrease in the voltage drop $I_1 r_b$. The charge accumulation and, hence, the decrease in voltage v_b depend on the mean lifetime τ .

A simplified analysis gives the following relation for a thick base:

$$v_b(\theta) = V_b(0) \left\{ 1 - \frac{L}{w} \ln \left[\frac{1}{2} \frac{I_1}{I_0} \left(\frac{\rho_{b0}}{\rho_i} \right)^2 + 1 \right] \operatorname{erf} \sqrt{\theta} \right\} \quad (3.41)$$

where $V_b(0) = I_1 r_{b0}$ is the initial voltage across the **nonmodulated** base resistance r_{b0} . We define the initial surge ΔV_b (see Fig. 3.18b) as the difference between $V_b(0)$ and $V_b(\infty)$:

$$\Delta V_b = V_b(0) \frac{L}{w} \ln \left[\frac{1}{2} \frac{I_1}{I_0} \left(\frac{\rho_{b0}}{\rho_i} \right)^2 + 1 \right] \quad (3.42)$$

For example, if $I_1/I_0 = 10^{10}$, $\rho_{b0}/\rho_i = 10^{-5}$ (at $\rho_{b0} = 2 \Omega \text{ cm}$), and $w/L = 2$, then $\Delta V_b \approx 0.2 V_b(0)$.

The *transition time* (at a level $V_b(\infty) + 0.1 \Delta V_b$) is found to be equal to

$$\theta_{t2} \approx 1.6 \quad (3.43)$$

As seen, $\theta_{t2} \gg \theta_{t1}$. Hence, it is safe to neglect the transient $v(\theta)$, considering that the stationary voltage, as given by Eq. (3.22), sets in **immediately** after arrival of the pulse I_1 .

We have assumed above that the length of pulse I_1 exceeds θ_{t2} . Where narrower pulses are involved, charges have no time to build up completely and the relative value of voltage surge diminishes; it also does so when the forward current I_1 decreases, as is apparent from Eq. (3.42).

3.3.5. Removal of excess carriers. After switching the diode from the forward to the reverse state, the charge stored up in the base cannot change in an instant. This is particularly obvious if we regard the diode as being shunted by the diffusion capacitance indicative of the presence of the above charge. Like any other capacitance, this capacitance recharges gradually under the action of the current flow. For the same reason the voltage across the junction does not change in an instant after switching; it then drops smoothly to zero and further grows, with its sign reversed, to the steady-state value e_2 (see Fig. 3.18b). The *storage time* for the charge in the base is the time interval between the moment of reverse switching and the moment at which the forward voltage v across the junction drops to zero¹.

If at the stage of charge removal the relation $v \ll e_2$ holds true, then the reverse current proves equal to the specified value:

$$-I_2 = \frac{-e_2 + v}{R} \approx \frac{-e_2}{R} = \text{constant}$$

¹ Analysis shows that at $v = 0$ a certain residual charge still remains in the base, but its removal takes place in the next stage of the transient process.

Solving transient diffusion equation (2.64) and considering relation (3.13a), it is possible to obtain the transient characteristic for forward voltage in the storage interval:

$$v(\theta) = \varphi_T \ln \left(\frac{I_1}{I_0} - \frac{I_1 + I_2}{I_0} \operatorname{erf} \sqrt{\bar{\theta}} + 1 \right) \quad (3.44)$$

Assuming $v(\theta) = 0$, we find the storage time in the implicit form

$$\operatorname{erf} \sqrt{\bar{\theta}_s} = \frac{I_1}{I_1 + I_2} \quad (3.45a)$$

Using the approximation $\operatorname{erf} \sqrt{\bar{\theta}} = \sqrt{1 - e^{-\theta}}$ yields the expression for storage time in the explicit form

$$\theta_s = -\ln \left[1 - \left(\frac{I_1}{I_1 + I_2} \right)^2 \right] \quad (3.45b)$$

For example, if $I_2 = I_1$, then $\theta_s \approx 0.3$. As the reverse current I_2 grows, the storage time decreases, which is quite natural since the large current sweeps out more speedily the charge from the base.

At the moment of switching, the diode current varies in magnitude from I_1 to $-I_2$, that is, by a value of $I_1 + I_2$. The voltage across the base layer correspondingly decreases stepwise by the value

$$\Delta V_b = (I_1 + I_2) r_b$$

The total voltage V_d drops by the same value since the component V does not change in switching. The drop ΔV_b is called *ohmic*.

A variant of charge removal is the case where the diode is not switched over but turned off, so that $e_2 = 0$ and $I_2 \approx 0$. Then, after the ohmic drop $\Delta V_b = I_1 r_b$, the voltage v_b falls nearly to zero and further stays invariable. Consequently, the voltage across the diode will be determined by the voltage v on the junction. This voltage, according to Eq. (3.44) at $I_2 = 0$, may be written as

$$v(\theta) = \varphi_T \ln \left[\frac{I_1}{I_0} (1 - \operatorname{erf} \sqrt{\bar{\theta}}) + 1 \right] \quad (3.46)$$

If we separate out the factor $(I_1/I_0 + 1)$ in the brackets of Eq. (3.46), take the logarithm of the found product and set $I_0 \ll I_1$, then expression (3.46) will assume the form

$$v(\theta) = V(0) + \varphi_T \ln (\operatorname{erfc} \sqrt{\bar{\theta}}) \quad (3.47a)$$

where $V(0)$ is the voltage across the junction at the moment of switching, and $\operatorname{erfc} \sqrt{\bar{\theta}}$ is a complementary error function (see p. 72).

Analysis shows that the initial and the finite region of the transient characteristic are of little significance. For the main region characterized by $\theta > 0.5$, the approximation $\operatorname{erf} \sqrt{\bar{\theta}} = 1 - (1/2) e^{-\theta}$ is applicable. This yields a linearly drooping characteristic repre-

mented by

$$v(\theta) = V(0) - \varphi_T \theta \quad (3.47b)$$

Setting $v(\theta) = 0$ and using Eq. (3.22) for $V(0)$, we determine the storage time from Eq. (3.47b)

$$\theta_s = \frac{V(0)}{\varphi_T} \approx \ln \frac{I_1}{I_0} \quad (3.48)$$

For example, if $I_1/I_0 = 10^{10}$, then $\theta_s \approx 20$. As seen, in the absence of reverse current the period of charge removal extends considerably [see the example relating to Eq. (3.45b)].

3.3.6. Recovery of reverse current. This portion of the transient characteristic is most complicated as regards its analysis. What involves the main difficulty is the need to deal with a rather intricate initial distribution of carriers in solving the diffusion equation.

If we apply quite an acceptable approximation

$$\Delta n(x, 0) = p_0 \frac{I_1}{I_2} (e^{-x/L} - e^{-x/l})$$

where $l = L/(1 + I_2/I_1)$, then, after rather awkward calculations and some omissions, the transient characteristic becomes comparatively simple in form:

$$v(\theta) = [I_1 \operatorname{erf} \sqrt{\bar{\theta}} - (I_1 + I_2)(1 - e^{(a^2-1)\bar{\theta}}) \operatorname{erfc} \sqrt{a^2 \bar{\theta}}] R \quad (3.49)$$

where $a = 1 + I_2/I_1$. Considering that in the recovery region the reverse current is related to voltage by the expression

$$i(\theta) = -I_2 - \frac{v(\theta)}{R}$$

It is easy to obtain the transient characteristic for reverse current¹:

$$i(\theta) = I_1 \operatorname{erfc} \sqrt{\bar{\theta}} - (I_1 + I_2) e^{(a^2-1)\bar{\theta}} \operatorname{erfc} \sqrt{a^2 \bar{\theta}} - I_0 \quad (3.50)$$

The recovery time of reverse current cannot be determined at $0.1I_2$ because the current at this level is many orders of magnitude higher than I_0 corresponding to the stationary reverse biased condition. To determine the *recovery time* θ_r , therefore, we assume quite a definite value of reverse current close to I_0 :

$$i(\theta_r) = -nI_0$$

Using in Eq. (3.50) the approximation $\operatorname{erf} \sqrt{\bar{\theta}} \approx 1 - (1/2)e^{-\bar{\theta}}$ and completing certain transformations, it is possible to obtain the

¹ The right side of the expression includes the reverse thermal current I_0 which is of no significance in formula (3.49), but here becomes an asymptotic value with $\theta \rightarrow \infty$.

time θ_r , expressed in the simple explicit form

$$\theta_r = \ln \frac{I_2/I_1}{2(n-1)} \quad (3.51)$$

Thus if $I_2/I_0 = 10^{10}$ and $n = 3$, then $\theta_r \approx 20$. This means that at $\tau = 1 \mu\text{s}$ the pulse tail (trailing edge) will have a duration of about $20 \mu\text{s}$, which heavily decreases the speed of response of the diode. In special pulse diodes the lifetime and correspondingly the recovery time can be 2 or 3 orders of magnitude smaller.

In distinction to the storage time, the recovery time is directly rather than inversely dependent on the initial reverse current I_2 . This is due to the fact that in defining the quantity θ_r , we have dealt with the definite level of final current, nI_0 ; consequently, the higher the upper, or initial, level of I_2 , the longer the time that is needed to reach the lower level. If the lower level were set at $0.1I_2$, then the recovery time θ_r would be independent of current and reach $\ln 5 \approx 1.6$.

3.4. Semiconductor-Metal Contacts

The first solid-state devices were *point-contact diodes* using a semiconductor-metal contact (junction) as the rectifying element. They owed their inception to the experimentally discovered effect—rectification of weak ac pulses on bringing a sharp metal point in close contact with crystals of some natural semiconducting minerals. The early devices of this type were certainly unreliable and had unstable and nonreproducible characteristics. But these devices served as the basis for the creation of more advanced point-contact diodes now in use and, above all, provided a great impetus to the development of modern transistor electronics¹. In integrated circuits, metal-semiconductor (silicon) contacts find two uses, either as nonrectifying, ohmic contacts to supply current to and draw it from integrated elements or as specific rectifying contacts, known as Schottky barrier diodes.

The structure and properties of metal-semiconductor contacts primarily depend on the relative positions of Fermi levels in either of the two layers. Fig. 3.19 shows the band diagrams for separated layers at the top and the band diagrams of respective junctions at the bottom after “bringing” the layers in contact and establishing equilibrium.

3.4.1. Rectifying junctions. Fig. 3.19a illustrates energy band diagrams for the layers in which $\varphi_{Fm} > \varphi_{Fp}$. Such a relationship

¹ The first transistors invented in 1948 were point-contact devices consisting of a slice of germanium and two sharp metal points resting on the slice surface. On applying voltages of opposite polarities to the points spaced 10 to 20 μm apart, current was found to flow from one circuit to the other.

indicates that the probability of occupation of an arbitrary energy level φ , if it lies in the conduction band of the semiconductor, is smaller than if it is in the metal. In other words, the occupancy of the conduction band in a semiconductor is lower than the occupancy of a similar energy region in a metal. Consequently as the layers "come" in contact, an amount of electrons will move from the metal to the p -type semiconductor. The additional electrons penetrating

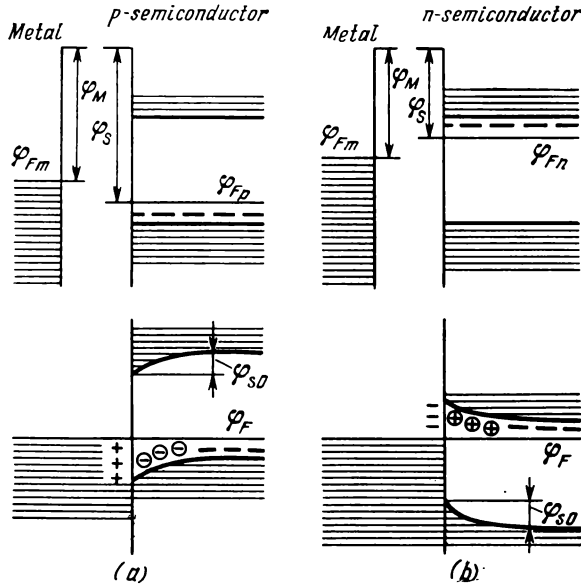


Fig. 3.19. Energy diagrams for rectifying metal-semiconductor contacts
(a) contact to p -semiconductor; (b) contact to n -semiconductor

into the surface layer of the semiconductor lead to intensive recombination. This decreases the amount of majority carriers—holes—and uncovers the uncompensated negative ions of acceptors in the boundary layer of a semiconductor. The electric field so built up impedes further inflow of electrons to the semiconductor, and brings about the Boltzmann equilibrium in the contact region. The energy levels here are seen to dip downwards.

In Fig. 3.19b are shown the band diagrams for the case where $\varphi_{Fm} < \varphi_{Fn}$; here, as the layers come in contact, electrons go from the n -type semiconductor to the metal, thereby uncovering the uncompensated positive ions of donors in the boundary layer of the semiconductor, which causes the bands to bend upwards.

The region of band bending (the region of space charges) in both cases had commonly a length of 0.1 or 0.2 μm , as determined by formula (2.53).

It is practically impossible to make a reliable junction between a metal and semiconductor by merely bringing them in touch with each other. Real contacts of this kind are at present produced by vacuum deposition of metal on a semiconductor slice.

An exchange of electrons between a metal and a semiconductor is generally defined by the difference between the *work functions* rather than by the difference between the "initial" Fermi levels. The work function of a solid is the energy required for the removal (thermal emission) of an electron out of the crystal. On the band diagrams, the work function is the energy "distance" between the level of a free electron outside the solid and the Fermi level. In Fig. 3.19, the work functions for the metal and semiconductor are respectively denoted by φ_M and φ_S . The difference between the work functions, $\varphi_{MS} = \varphi_M - \varphi_S$, expressed in volts, is termed the *contact-potential difference*.

Depending on the relation between the work functions φ_M and φ_S , electrons can pass to one layer or the other. If $\varphi_M < \varphi_S$ (that is, $\varphi_{MS} < 0$), electrons transfer from the metal into the semiconductor (see Fig. 3.19a), and if $\varphi_M > \varphi_S$ (that is, $\varphi_{MS} > 0$), they move from the semiconductor to the metal (see Fig. 3.19b). Such a criterion is more illustrative than the one used at the beginning of the Section, all the more so, because the contact-potential differences for standard combinations of metals and semiconductors are listed in the specialist literature.

The amount by which the energy bands are bent near the surface (see Fig. 3.19) is described by the equilibrium surface potential φ_{s0} (see p. 59). If we ignore the role of surface states, the quantity φ_{s0} will be equal to the contact potential difference φ_{MS} .

Both contacts shown in Fig. 3.19 feature **depletion** layers in the contact region of the semiconductor. Here the concentration of majority carriers is smaller than the equilibrium concentration kept stable far from the contact. Consequently, this *contact region has an increased resistivity and therefore determines the resistance of the entire system*. Such a situation is analogous to the one which we have noted in dealing with *pn* junctions on p. 79.

The potential barrier in the contact layer is called the *Schottky barrier*. Its height φ_{s0} is an analog of the quantity $\Delta\varphi_0$ in the *pn* junction. The potential φ_s and respectively the contact layer resistance will vary with the polarity of externally applied voltage.

Thus if the metal is at a positive voltage with respect to the semiconductor, the potential barrier at the contact shown in Fig. 3.19a grows, so that the contact layer will become yet more depleted and thus will have an increased resistance in comparison with the resi-

stance typical for the equilibrium state. So, the voltage of this polarity at the given contact is **reverse**. At the contact displayed in Fig. 3.19b, the polarity of applied voltage being the same, the potential barrier will decrease, with the result that the contact layer will get enriched with the majority carriers (electrons) and its resistance will be smaller than the resistance in the equilibrium state. The voltage of this polarity for the contact in question is **forward**.

Thus the contacts in Fig. 3.19 display rectifying properties and can serve as the basis for diodes. The diodes using the Schottky barrier are known as *Schottky barrier diodes*.

A variant of rectifying contacts is the contact which features an inversion layer in the semiconductor at the semiconductor-metal boundary, that is, the layer with the opposite type of conductivity. In Subsec. 2.7.4 we have made mention of the possibility for the formation of such layers in MIS systems. Fig. 3.20 shows the band diagram of the contact having an inversion *p*-layer. This case is specific to heavily bent bands, that is, to large contact potential differences, φ_{MS} , when the electrostatic potential level traverses the Fermi level near the boundary. The thickness

of an inversion layer, as noted earlier, does not exceed 1 or 2 nm.

The contact having an inversion layer is on the whole similar to the *pn* junction, since there are eventually two "contacting" layers of the *p*- and the *n*-type. However, *injection in such a structure is absent*. Note that if the band bending is still larger than that depicted in Fig. 3.20, the Fermi level will cross the level of the "top" of the valence band. In this case, a portion of the inversion layer adjacent to the boundary will convert to a degenerate semiconductor—**semimetal**.

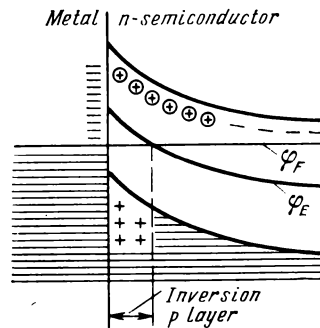


Fig. 3.20. Band diagram for a contact at which an inversion layer builds up

3.4.2. Schottky diodes. An important feature of Schottky diodes that distinguishes them from *pn* junctions is the *absence of injection of minority carriers*. These diodes *operate*, so to say, *entirely on majority carriers*. From this it follows that Schottky diodes are *free from diffusion capacitance* associated with the storage of minority carriers in and their sweeping out of the base (see p. 90). This effectively raises the response of the diodes with the change of currents and voltages, including the switching from the forward to the reverse condition and vice versa. The switching time here only depends on the *barrier capacitance*, and can be as small as tenths and hundredths

of a nanosecond in diodes of small area. The corresponding working frequencies at which these diodes operate lie in the range from 3 to 15 GHz.

Another equally important feature of Schottky diodes is a *substantially lower forward voltage* in comparison with the voltage at the *pn* junction. This is attributed to the fact that the *I-V* characteristic of Schottky diodes is described by the same classical formula (3.16) derived for *pn* junctions, but the thermal currents here are much higher since the diffusion rate D/L typical of the *pn* junction [see Eq. (3.18)] is replaced by the mean thermal velocity of carriers, v_T . The latter quantity exceeds D/L by approximately three orders of magnitude. From Eq. (3.22) it then follows that the forward voltage at Schottky diodes is about 0.2 V below that at the *pn* junction. The typical values of forward voltage for Schottky diodes lie in the neighborhood of 0.4 V. As for reverse currents, these may range from 10^{-11} to 10^{-12} A depending on the junction area; that is, they lie close to real reverse currents in silicon *pn* junctions, attributable to thermally generated current (see p. 92).

One more feature of Schottky diodes is that their *I-V* characteristic **strictly** obeys exponential equation (3.16) over a very wide range of currents, to the extent of a few decades, for example from 10^{-12} to 10^{-4} A. Hence, it is possible to use Schottky diodes as *precision logarithmic elements* in compliance with relation (3.22). This feature is also due to the absence of injection: in *pn* junctions, the presence of injected excess carriers causes a change (modulation) in base conductivity, which affects the shape of the *I-V* curve (see p. 91).

Reliable Schottky barriers are produced in silicon when in contact with such metals as molybdenum, nichrome, gold, platinum (more precisely, the alloy of platinum with silicon, that is, platinum silicide), and also aluminum, which is the basic material used for metallization in ICs. That Schottky barriers have come to be popular only quite recently, starting from the 1970s, though the theory on the barrier effect goes back over 50 years, is due to the following reasons. First, to obtain a quality barrier, it is necessary to provide a "fused-in" contact rather than a pressed-on metallic contact to a semiconductor. It is not until the advent of the technique of vacuum deposition of films that such a contact has become realizable. Second, it is necessary that the base of a diode in particular should have a small resistance at a sufficiently high breakdown voltage, though, as we know, the breakdown voltage falls off with decreasing base resistivity (see Subsec. 3.2.7). It has become possible to solve

¹ The rate of diffusion was defined as L/τ on p. 70. Multiplying the numerator and denominator by L and substituting $L^2 = D\tau$ [see Eq. (2.66)] yields $L/\tau = D/L$. As regards Schottky diodes, the replacement of diffusion rate by thermal velocity is quite natural since these diodes do not exhibit diffusion due to injection of minority carriers.

the problem only after the development of epitaxial technology (see Sec. 6.3), which enables the growth of a high-resistivity working film on a low-resistivity substrate.

3.4.3. Nonrectifying contacts. Let the inequality $\varphi_{MS} > 0$ hold for the contact of a metal with a p -type semiconductor, and the inequality $\varphi_{MS} < 0$ for the contact between a metal and an n -type semiconductor (Fig. 3.21). As we already know, in the former system

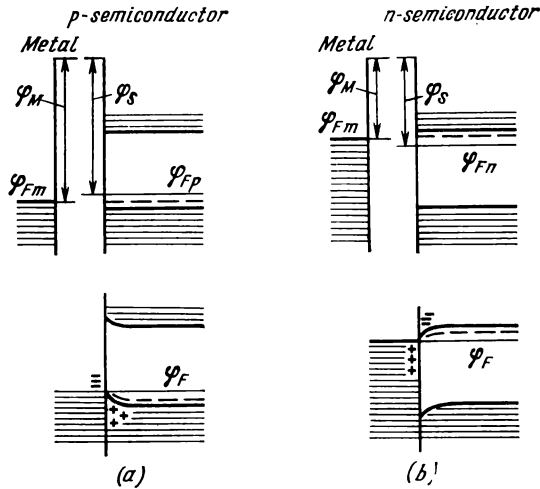


Fig. 3.21. Energy diagrams for nonrectifying metal-semiconductor contacts (a) contact to p -semiconductor; (b) contact to n -semiconductor

electrons will pass from the semiconductor to the metal, causing the band edges to bend upwards, while in the latter system electrons will move from the metal to the semiconductor, so that the band edges will dip towards the surface. In such contacts a semiconductor becomes rich in **majority** carriers near the interface, which form **enriched** layers whose depth is definable by the Debye length [see Eq. (2.52)] and is as small as hundredths of a micrometer. As clear from Fig. 3.21, we have assumed contact potential differences to be very small, therefore the band bending is small and the semiconductors remain nondegenerate. If we set $\varphi_{MS} = 0.1$ or 0.2 V and over, then the amount of band bending will be much greater, and the Fermi level will cross a corresponding energy band near the boundary. In this region the semiconductor degenerates to become a semimetal of extremely low resistivity.

Irrespective of whether or not the semiconductor becomes degenerate, the presence of the enriched layer is indicative of the fact that *the resistance of the system as a whole is determined by the semiconductor neutral layer* and, hence, is independent of either the value or the po-

larity of applied voltage. Such nonrectifying combinations of a metal with a semiconductor are called *ohmic contacts*.

Ohmic contacts are made at the points of connection of leads to semiconductor layers. The fabrication of ohmic contacts is the task of no less significance than that of producing rectifying junctions. An important property of an ohmic contact, apart from its two-way conduction is an extremely short lifetime of excess carriers. In the analysis of semiconductor devices, therefore, it is usual to assume that *the concentration of excess carriers at an ohmic contact is equal to zero*.

The most popular metal used in modern microelectronics for ohmic contacts is aluminum, which is deposited by spraying on the surface of silicon and then "fired on" (that is, fused, or alloyed) to a small depth at an elevated temperature. If silicon is of the p type, aluminum (being an acceptor) additionally dopes the surface layer during alloying and thus adds to the conductance of the ohmic contact.

If silicon is of the n type, then in the fusion process the acceptor atoms of aluminum may overcompensate the host donor atoms, and the layer of silicon just under the surface will become of the p type. The resultant contact is a parasitic pn junction. The probability that this can be the case is higher with a lower concentration of donors, that is, at a high resistivity of the n -type silicon. To avoid such an occurrence, the surface of n -silicon in the region of contact is additionally doped with donors to form an n^+ layer where the overcompensation by aluminum atoms cannot take place (see p. 201).

3.5. Semiconductor-Insulator Interface

In Ch. 2 we have pointed out more than once that the surface layer of a semiconductor is a specific region with its properties noticeably different from those of the bulk. The surface layer has a particular crystal structure, contains particular (adsorbed) impurities, and exhibits characteristic energy levels. Consequently, the surface layer inherently differs from the bulk in the carrier mobility, lifetime, and other electrophysical parameters.

3.5.1. General characteristic of the Si-SiO₂ interface. Needless to say that the properties of a medium with which a semiconductor comes in contact exert an influence on the properties of its surface layer. An example may be the junctions (contacts) between semiconductors and metals discussed in the preceding Section. As has been revealed earlier, a metal present on the surface of a semiconductor tends to form depletion or enriched layers. Analogous processes occur at the boundary between a semiconductor and a dielectric.

Of particular interest is the boundary between silicon and silicon

dioxide since the surface of all modern semiconductor ICs is protected with an oxide layer (see Fig. 1.3). Besides, in MOS structures (see Fig. 2.20) based on silicon it is the SiO_2 layer that generally serves as an insulator. Therefore, in dealing below with the semiconductor-insulator interface, we shall imply that under discussion is the Si- SiO_2 structure which has the highest practical significance.

The main feature of oxide layers (films) used in ICs is that they *always contain donor impurities*, of which the most common are sodium, potassium, and hydrogen. These elements are present in standard solutions intended for the treatment of silicon slices and

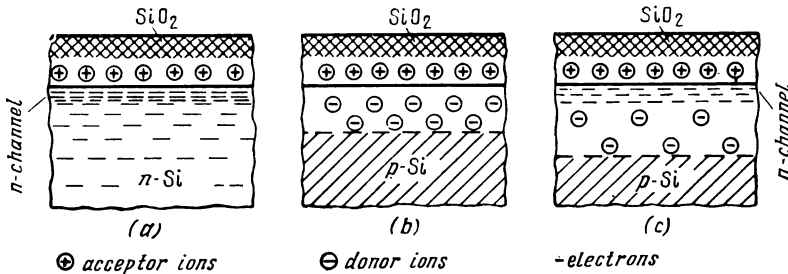


Fig. 3.22. Si- SiO_2 interface region structure

(a) enriched layer; (b) depletion layer; (c) depletion layer with inversion channel

also in glass and quartz that go into the production of vessels and auxiliaries employed in the manufacturing processes (see Ch. 6).

Practical experience shows¹ that donor impurities typical of a SiO_2 film are concentrated near the surface of silicon. In the SiO_2 film, therefore, where it comes in contact with silicon, a thin layer of positive donor atoms appears, because electrons leave the donors and escape into the surface layer of silicon. The effect of such an escape of electrons depends both on the type of semiconductor conductivity and on the concentration of donor impurities in the insulator. Since the donor atoms lie in a very thin layer of the insulator, the volume concentration (cm^{-3}) proves an inconvenient parameter, and so the use is made of the surface concentration (cm^{-2}). The typical values of the surface concentration of donors in silica, $N_{d\text{SiO}_2}$, range from 0.5×10^{12} to $2.0 \times 10^{12} \text{ cm}^{-2}$.

If Si has the n -type conductivity, then the electrons that have left the oxide for the semiconductor will enrich the Si surface layer with majority carriers, thereby producing what is called an n channel (Fig. 3.22a). If Si is of the p type, then the electrons that have come

¹ The initial impurity distribution in SiO_2 is nearly uniform. Then in the course of time, high temperature specific to basic production processes (see Ch. 6) causes the impurities to diffuse toward the Si- SiO_2 boundary because silicon lacks such impurities.

from the oxide to the semiconductor either deplete the surface layer of carriers, uncovering the negative acceptor ions (Fig. 3.22b), or form, along with the depletion layer, a thin inversion n layer (Fig. 3.22c).

3.5.2. Effect of the Si-SiO₂ interface on the parameters of pn junctions. The n channels and also depletion and inversion layers formed at the contact between silicon and the SiO₂ film have a definite and, sometimes, considerable effect on the operation of semiconductor devices and integrated elements. As to MOS structures, this effect

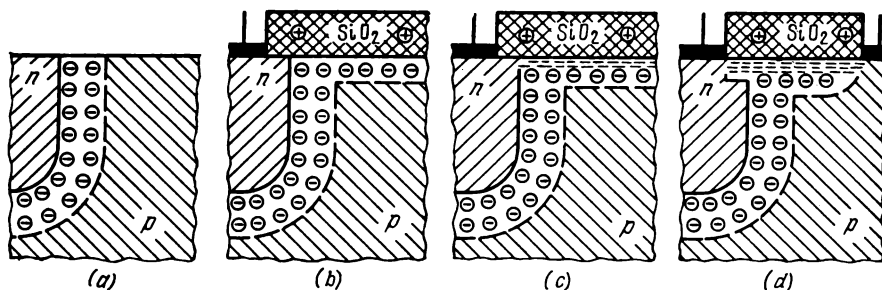


Fig. 3.23. Structure of planar pn junctions near the surface

(a) in the absence of donor impurities in oxide; (b), (c) and (d) in the presence of donor impurities

is discussed in Subsec. 5.2.1. Consider now the Si-SiO₂ interface effect on the operation of planar pn junctions whose external regions extend as far as this interface (see Fig. 1.3).

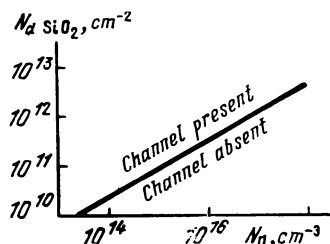
Figure 3.23a shows a "vertical" (side) section of the pn junction in the absence of donors in the SiO₂ film. Where donors are present, the surface depletion layer formed in p -Si (Fig. 3.22b) merges into the initial "vertical" depletion layer (Fig. 3.23b). The total area and volume of the depletion layer thus grow, entailing an increase in the thermally generated current according to Eq. (3.26), which is the main component of reverse current in silicon pn junctions. What favors still further growth of reverse current is a short lifetime in the surface layer, as evident from Eq. (3.26).

If, along with the depletion layer, an inversion n channel forms in p -silicon (Fig. 3.22c), then this channel combines with the n layer of the junction and extends it, as it were, along the surface (Fig. 3.23c). As in the preceding case, the resultant area of the depletion layer increases. But now this layer is separated from the surface by the conducting n channel. Thus the depletion layers' surface portion noted for a short lifetime does not contribute to the thermally generated current, and so this current proves smaller in value than in the

preceding case. Despite this, n channels on the whole play rather a negative part.

Indeed, an n channel located in the p layer is able to form a conducting bridge between ohmic contacts of the n - and p -layers and thus *short out* the pn junction (Fig. 3.23d). Besides, even in the absence of the short, the horizontal (surface) portion of the depletion layer has a smaller thickness than the vertical portion, so the former stands up to a lower breakdown voltage. Last, the electrons forming

Fig. 3.24. Relationship between the donor concentration in oxide and acceptor concentration in silicon, which determines the probability of inversion channel formation



the n channel readily fall into traps lying in the oxide and then come back to the channel, thereby causing fluctuations of the current through the pn junction. This shows up as intrinsic noise of an increased level.

Whether the inversion layer is present or not depends both on the concentration of donors in the oxide film and on the concentration of acceptors in the junction p layer. This dependence is represented by the graph in Fig. 3.24. Thus, if the surface concentration of donors in the oxide is $N_{d\text{SiO}_2} = 10^{12} \text{ cm}^{-2}$, the channel can form when the acceptor concentration $N_a < 6 \times 10^{16} \text{ cm}^{-3}$. The lower the acceptor concentration, that is, the higher the resistivity of the p layer, the higher the probability that an n channel can appear.

As regards junction n layers, the donor impurities located in the oxide film, form an excess of electrons near the surface, that is, **enriched layers** (see Fig. 3.22a). These layers play an important part in MOS structures (see Sec. 5.2), but are of minor significance in pn junctions.

4

4.1. General

Transistors are semiconductor amplifying devices, that is, the devices capable of amplifying electrical power¹. There is a great variety of transistors that differ in design and structure, but by the principle of action they fall into two basic classes, covering respectively *bipolar* (junction) and *unipolar* (field-effect) transistors.

The underlying mechanism of bipolar transistor operation is the injection of minority carriers, for which reason *pn* junctions are an inherent part of bipolar transistors. The term bipolar reflects the fact that **both** types of charge carrier, electrons and holes, participate in the operation of transistors: the injection of **minority** carriers is attended with the compensation of their charge by **majority** carriers (see Subsec. 2.8.3).

The purpose of this chapter is to study the physical processes in a bipolar transistor and also analyze its basic characteristics and parameters. Unipolar transistors are discussed in the next chapter.

4.2. Transistor Action

The bipolar transistor is a combination of two **interacting** *pn* junctions connected in an opposite polarity relationship. The interaction between the junctions is due to a fairly close location of one junction with respect to the other, at a distance that is smaller than the carrier diffusion length.

4.2.1. Transistor structure. In real transistors, one *pn* junction differs substantially from the other, as is clear from Fig. 4.1a: the n_1p junction has by far a smaller area than the n_2p junction. Besides, in most transistors one of the extreme layers (namely, the small-area layer n_1) is doped much more heavily than the other, n_2 . In this respect, the transistor is an asymmetric device.

The names of the extreme layers reflect the asymmetry of a transistor: the highly doped small-area layer n_1 is called the *emitter*, and the large-area layer n_2 is called the *collector*. The junctions n_1p

¹ It is exactly the gain in power rather than in voltage or current that makes a criterion by which one can place a device into the class of amplifiers. For example, a transformer amplifies voltage at the expense of current or current at the expense of voltage, but does not amplify power, and therefore it cannot belong to the class of amplifying devices.

and n_2p between these layers and the middle p -type layer, known as the *base*, are respectively termed the emitter and collector junctions. The meaning of these terms is explained below.

Each of the pn junctions of a transistor has a *bottom* portion and *side* portions.

The working zone, or what is called the *active zone* of a transistor, is the region located under the bottom portion of the emitter junction (the unhatched area on Fig. 4.1a). The hatched portions of the

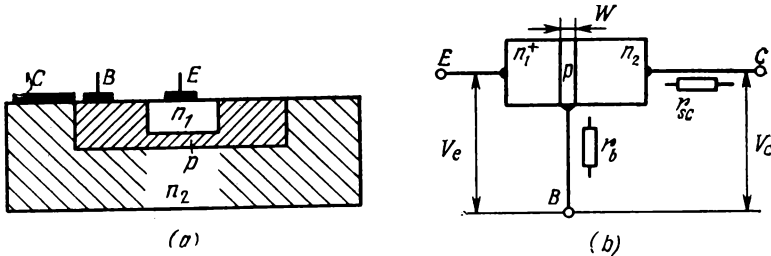


Fig. 4.1. Structure of a bipolar transistor
(a) real; (b) ideal, without passive regions

structure are *passive* and, in a way, parasitic, but from design and technological considerations they make an inherent part of the transistor structure (see Sec. 6.9). The passive regions may be represented to a first approximation by the resistors connected to the working layers of the base and collector.

Figure 4.1b shows the transistor's active portion in its horizontal position and also the points of connection of resistors, r_b and r_{sc} , representative of passive regions, the highly doped layer (emitter) being designated as n^+ . The structure presented in Fig. 4.1b serves as the basis for the analysis of transistors.

The emitter junction interacts with the collector junction owing to a small thickness, or width, w of the base. In modern transistors, the base width does not exceed $1\text{ }\mu\text{m}$, whereas the diffusion length L ranges from 5 to $10\text{ }\mu\text{m}$.

The basic properties of a transistor are determined by the processes taking place in the base. If the base is homogeneous (uniform), then the transport of carriers in it is due purely to diffusion. If the base is inhomogeneous (nonuniform), then, as is known (see Subsec. 2.4.7), it exhibits an internal (built-in) electric field, so that the motion of carriers in the base is of the combined type; namely, the diffusion and drift of carriers occur in combination. Transistors with a homogeneous base are known as *diffusion* (drift-free) transistors, and those with an inhomogeneous base as *drift* transistors. The latter find at present most extensive use in integrated circuits.

The transistor shown in Fig. 4.1 that consists of the n -type emitter and collector layers and the p -type base layer sandwiched inbetween is known as an npn transistor. Since npn transistors play a leading role in microelectronics, we shall use mainly these structures in the subsequent analysis. Transistors with the p -type emitter and collector regions and the n -type base region, known as pnp transistors, also find applications. By the principle of action, the pnp transistor does not differ from the npn type, but its working voltages are of the opposite polarity and it also shows a number of quantitative distinctions.

4.2.2. Modes of operation. In its normal mode of operation, the transistor has its emitter junction **forward biased** and the collector **reverse biased**. The emitter emits the electrons through the base, which cross it freely hardly sustaining any loss by recombination (since the base region is narrow), and enter into the collector region maintained at the positive potential (see the band diagram of a transistor in Fig. 4.2). Thus, in the normal mode of operation the collector *collects* the minority carriers being injected into the base. Hence, the name the collector.

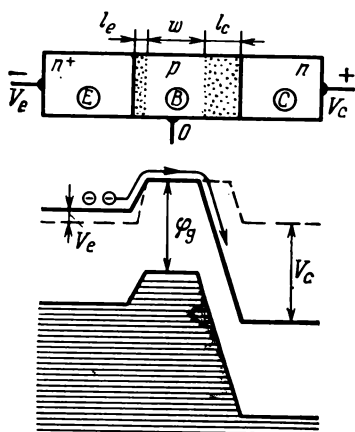


Fig. 4.2. Band diagram for a transistor in normal mode of operation

It is clear that with positive polarity, collector is capable of collecting only **electrons**. It is therefore important that the emitter current should largely consist of **electrons**. That is why the emitter is doped far more heavily than the base to make the emitter junction **one-sided** (see Subsec. 3.2.3). As for the collector junction,

this is commonly one-sided in diffusion transistors, but as a rule almost symmetric in drift transistors.

In the normal operating condition, the collector and emitter currents are almost equal, the small difference between the two being a base current. The base current makes up for the loss of **majority** carriers (holes) due to recombination, which inevitably occurs even if the base is very narrow, and also due to the injection of holes from the base to the emitter.

The resistance of the reverse-biased collector junction is very high, a few megohms and over. This makes it possible to insert a rather large load resistance in the collector circuit without changing the collector current and thereby obtain a substantial power in the

load circuit. The resistance of the forward-biased emitter junction is, on the contrary, very small; for example, at a current of 1 mA, this resistance is merely $25\ \Omega$ [see Eq. (3.27)]. Therefore, with the currents in the emitter and collector being almost the same, the power delivered to the emitter circuit may be by far smaller than the power derived in the load circuit. Hence, the transistor is able to amplify power, that is, to operate as an amplifier.

Despite the asymmetry of a transistor, it is possible to interchange the roles of the emitter and collector by applying the forward bias to the collector junction and the reverse bias to the emitter junction. With this reverse switching, the transistor is said to be in the *inverted mode* or in the inverse region of operation. The transfer of current with the transistor operating in the inverse region is much worse than when it operates in the normal mode. The reasons are as follows. First, the electron component of the collector current is small due to a light doping of the collector. Second, the area of the real collector is much greater than the area of the emitter (see Fig. 4.1a), so that only a small part of electrons injected from the collector get into the emitter.

A particular mode of transistor operation is the *mode of double injection*, or *saturation* (the origin of the latter term is explained in Subsec. 8.3.1). In this mode of transistor operation, the collector and emitter junctions are both **forward** biased; they inject carriers into the base, and at the same time each collects the carriers emitted by the other.

So far we have set voltages at the emitter and collector with respect to the base (see Fig. 4.1b). In this case the connection of the transistor into circuits to give an input and an output is called the *common-base* (CB) *connection* (the transistor is said to be connected in a common-base configuration or in a common-base circuit). Let us recall that one cannot practically specify the forward voltage at the *pn* junction and it is usual practice to set the forward current (see Subsec. 3.2.5). Hence, *the specified parameter for common-base transistors is the emitter current*.

The CB configuration helps reveal well the physical processes in the transistor; it also has some other attractive features. But since it does not permit current gain and has a low input resistance equal to the emitter junction resistance, this transistor configuration proves unfavourable for use in many applications. The transistor circuit connection that plays the main part in transistor engineering is the *common-emitter* (CE) *connection* typical of which is *the specified value of base current*. Fig. 4.3 shows both these configurations with the notation and circuit symbols adopted for the *npn* transistor.

Similar connections for the *pnp* transistor appear in Fig. 4.4. As mentioned earlier, characteristic of this transistor is the reverse

polarity of working voltages and, thus, the opposite direction of working currents.

In microelectronics, *pnp* transistors do not function as independent elements, that is, do not find use as **alternatives** to *npn* transistors

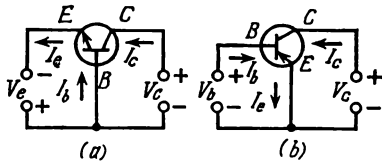


Fig. 4.3. *NPN* transistor in CB connection (a) and CE connection (b)

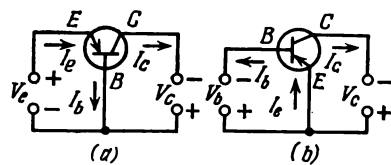


Fig. 4.4. *PNP* transistor in CB connection (a) and CE connection (b)

in the circuits of the same class. But they have opened the way for *complementary* integrated circuits using *npn* and *pnp* transistors in **combination**. In a number of cases, such a combination of complementary *npn* and *pnp* transistors simplifies the structure and optimizes the parameters of certain circuits.

4.3. Carrier Distribution

To calculate currents, voltages, and excess charges in a transistor, it is necessary to know the distribution of excess concentrations, that is, functions $\Delta n(x)$ and $\Delta p(x)$. We shall consider these functions for the main element in IC—*npn* transistor shown in Fig. 4.1b.

Functions Δn and Δp coincide by virtue of quasineutrality [see Subsec. 2.5.1 and Eq. (2.29)]. Therefore, the expressions below are given only for the excess concentrations of *minority* carriers.

4.3.1. Diffusion transistor in normal operation. Under the steady-state conditions, the concentration of carriers injected into the base is described by diffusion equation (2.65), the general solution of which takes the form as given by (2.67). The coefficients A_1 and A_2 are determined from the boundary conditions which apply to the emitter and collector boundaries.

In writing the boundary conditions for the transistor in normal operation we assume that the specified values are the value of reverse voltage U_c at the collector junction and the value of forward current I_e (more exactly, its electron component I_{en}) at the emitter junction.

Setting $|U_c| > 3\phi_T$, we find from Eq. (3.13a): $\Delta n = -n_0$. Since the equilibrium concentration of electrons in the *p*-type base is very small, we ignore the quantity n_0 and write the first boundary

condition in the form

$$\Delta n_b(w) = 0 \quad (4.1a)$$

Substituting the current density I_{en}/S in the left side of Eq. (2.57a), we can readily find the electron concentration gradient and thus the second boundary condition:

$$\left. \frac{d(\Delta n_b)}{dx} \right|_{x=0} = -\frac{I_{en}}{qD_bS} \quad (4.1b)$$

where D_b is the diffusion coefficient for minority carriers in the base. The minus sign ahead of the right member of Eq. (4.1b) reflects the fact that the forward (**positive**) emitter current in *npn* transistors

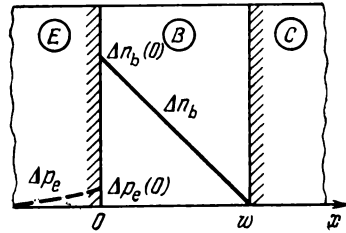


Fig. 4.5. Electron distribution in the base of a diffusion *npn* transistor

(see Fig. 4.3a) is indicative of the injection of electrons into the base, in which case the gradient of electron concentration must be **negative** (see Fig. 2.25).

Using boundary conditions (4.1), it is possible to determine the coefficients A_1 and A_2 in general expression (2.67) and then reduce it to the form

$$\Delta n_b(x) = I_{en} \frac{L}{qD_bS} \frac{\sinh[(w-x)/L]}{\cosh(w/L)} \quad (4.2)$$

Since the inequality $w \ll L$ holds for transistors, the above expression can be simplified if we resort to the relationships $\sinh(z) \approx z$ and $\cosh(z) \approx 1$ valid for small arguments:

$$\Delta n_b(x) = I_{en} \frac{w}{qD_bS} \left(1 - \frac{x}{w}\right) \quad (4.3)$$

As seen, *diffusion transistors with a uniform base show an almost linear distribution of excess carriers* (Fig. 4.5).

Integrating the function $\Delta n_b(x)$ over the range from 0 to w and multiplying the integral by the area S and elementary charge q yields the expression for the excess charge in the base:

$$\Delta Q_b = I_{en} (w^2/2D_b) \quad (4.4)$$

As clear from the formula, the excess charge is proportional to the emitter current and decreases at the given current as the base thickness becomes narrower.

We now turn to the distribution of holes injected from the base into the emitter. Because the emitter layer is much thicker than the base region, the former has w_e much greater than L_e , where w_e is the emitter layer thickness and L_e is the diffusion length of carriers (holes) in the emitter layer. Given this inequality, the excess carrier distribution will be the same as in the base of infinite thickness, that is, exponential in form (see Fig. 2.25). Replacing in Eq. (4.2) the electron component I_{en} by the hole component I_{ep} , the concentration Δn_b by Δp_e , and assuming $w \rightarrow \infty$, we get

$$\Delta p_e(x) = I_{ep} \frac{L_e}{qD_e S} e^{-x/L_e} \quad (4.5)$$

The distance x here is counted off from the emitter boundary into the emitter bulk (dash line in Fig. 4.5). The relation between the boundary concentrations $\Delta n_b(0)$ and $\Delta p_e(0)$ is determined by Eq. (3.14).

Let us integrate Eq. (4.5) between the limits $x = 0$ and $x = \infty$ and multiply the result by S and q . Considering Eq. (2.66), we then obtain the excess charge in the emitter

$$\Delta Q_e = I_{ep} \tau_e \quad (4.6)$$

where τ_e is the lifetime of minority carriers in the emitter layer.

4.3.2. Drift transistor in normal operation. The concentration of carriers in a nonuniform base is described by continuity equation (2.79). For the first boundary condition we can use expression (4.1a):

$$\Delta n_b(w) = 0 \quad (4.7a)$$

In writing the expression for the second boundary condition, we take account of the fact that the current I_{en} in the drift transistor is the sum of the diffusion and drift components. We now sum up the current densities given by (2.56a) and (2.57a) and equate the result to I_{en}/S . Next substituting the field strength E given by (2.77) and using (2.58), it is easy to reduce the second boundary condition to the form

$$\left. \frac{d(\Delta n_b)}{dx} \right|_{x=0} - \frac{\Delta n_b(0)}{L_N} = - \frac{I_{en}}{qD_b S} \quad (4.7b)$$

where L_N is the average depth of base doping [see Eq. (2.75)]. As for the minus sign placed in front of the right-hand member of Eq. (4.7b), the explanation is the same as for Eq. (4.1b).

In its general form, the expression for $\Delta n_b(x)$ determined under the boundary conditions as given by Eqs. (4.7) proves too complex

and nonillustrative. It becomes much simpler at $w \ll L$. In this case

$$\Delta n_b(x) = I_{en} \frac{w}{qD_b S} \frac{[1 - \exp[-2\eta(1 - \frac{x}{w})]]}{2\eta} \quad (4.8)$$

where $\eta = w/2L_N$.

The quantity η may be presented in a more descriptive form if we use Eq. (2.75). Substituting $x = w$ and taking the logarithms of both sides gives

$$\eta = 1/2 \ln [N_b(0)/N_b(w)] \quad (4.9)$$

where $N_b(0)$ and $N_b(w)$ are impurity concentrations at the emitter and collector boundaries of the base. The greater the difference between the impurity concentrations, the greater the value of η , for which reason this quantity is called the *coefficient of base inhomogeneity*¹. The typical values of η used in practice lie between 2 and 3.

Figure 4.6 shows the carrier distribution given by Eq. (4.8), the graphs being plotted to a relative scale. The unit of scale taken here is the boundary concentration of carriers in a diffusion transistor [see (4.3)]:

$$\Delta n_D(0) = I_{en} (w/qD_b S)$$

where the subscript D denotes the diffusion nature of the carrier motion. It is seen from Fig. 4.6 that as η grows, the carrier distribution in the base of a drift transistor becomes increasingly non-linear. Given the same values of emitter current, the drift transistor has a considerably **smaller** excess concentration than the diffusion transistor.

Integrating the function (4.8) and multiplying the result by S and q , we get the excess charge in the base:

$$\Delta Q_b = I_{en} \frac{w^2}{2D_b} \left(\frac{2\eta - 1 + e^{-2\eta}}{2\eta^2} \right) \quad (4.10a)$$

The composite function enclosed in the brackets approximates well to a simple function $(\eta + 1)^{-1}$, and thus the excess charge assumes

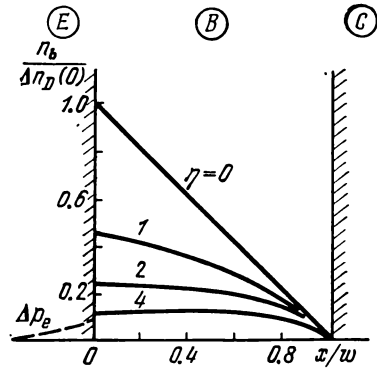


Fig. 4.6. Electron distribution in the base of a drift transistor

¹ In Subsec. 2.8.4 we considered the motion of carriers in an inhomogeneous semiconductor, but for other boundary conditions: the layer thickness was assumed infinitely large and the carrier concentration at the emitter boundary specified; the latter condition is equivalent to the specified voltage at the junction, as seen from (3.13a). The relation between the field coefficient θ [see (2.78)] and the base inhomogeneity coefficient η has the form $\eta = \theta (w/L)$.

the form

$$\Delta Q_b = I_{en} [w^2/2 (\eta + 1) D_b] \quad (4.10b)$$

The distribution of holes injected from the base into emitter is shown in Fig. 4.6 by a dash line. This distribution and the corresponding excess charge can be evaluated by Eqs. (4.5) and (4.6).

4.3.3. Transistors in inverse and double injection operation. In the inverse region of operation the distribution of carriers both in the

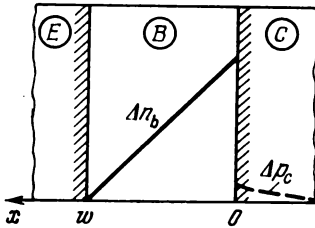


Fig. 4.7. Electron distribution in the base of a diffusion transistor in inverse mode of operation

base and collector of the **diffusion** transistor having a one-sided collector junction, turns out to be practically the same as in the normal operating mode (see Figs. 4.5 and 4.7).

As regards **drift transistors**, the inverted mode of operation brings about a qualitatively different distribution of carriers in the base. The reason is that the base field proves **retarding** rather than accelerating for the electrons injected from the collector. Changing the sign of η in (4.8), we obtain the excess carrier distribution in the base:

$$\Delta n_b(x) = I_{cn} \frac{w}{qD_bS} \frac{\exp \left[2\eta \left(1 - \frac{x}{w} \right) \right] - 1}{2\eta} \quad (4.11)$$

where I_{cn} is the electron component of collector current, the coordinate x being counted off from the collector to the emitter. The corresponding curves of carrier distribution appear in Fig. 4.8. Comparing Fig. 4.8 with Fig. 4.6, we see that the excess concentrations in the drift transistor operating in the inverse region are much higher than they are in the diffusion transistor. The excess charges are likewise larger in the drift transistor. Reversing the sign of η in (4.10a) we obtain the excess charge in the base:

$$\Delta Q_b = I_{cn} \frac{w^2}{2D_b} \frac{e^{2\eta} - 2\eta - 1}{2\eta^2}$$

For the excess concentrations and charges in the collector, Eqs. (4.5) and (4.6) can hold if we change the subscripts e for c . Since the collector junction in drift transistors is almost symmetric, the boundary concentrations of electrons and holes, according to (3.14), are

almost equal (see the dash line in Fig. 4.8). The electron and hole components of the collector current and also the excess charges in the collector and in the base must then be comparable in value.

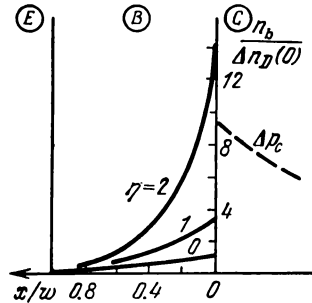


Fig. 4.8. Electron distribution in the base of a drift transistor in inverse mode of operation

With the transistor operating in the double injection mode in which both the emitter and collector junctions are forward biased, the distribution of excess carriers in the base may approximately

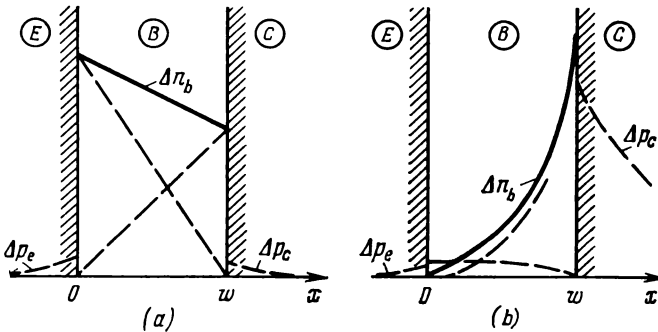


Fig. 4.9. Electron distribution in the base of a transistor operating in the double injection mode (dash lines are the components representative of the normal and the inverse mode of operation)

(a) diffusion transistor; (b) drift transistor

be estimated by **summing up** the distributions specific to the normal and the inverted mode of operation (Fig. 4.9). In diffusion transistors (Fig. 4.9a) the distribution is trapezoidal in form, and the excess charge is much larger than in normal operation. As to drift transistors (Fig. 4.9b), the resultant carrier distribution and excess charge are close in value to the carrier distribution and excess charge characteristic for inverse operation.

The total excess charge stored in all the three layers of a transistor can be expressed through the base current. Indeed, the current I_b is nothing but the rate of growth of the positive charge in the base. Under the steady-state conditions the base neutrality will keep stable if the rate of positive charge **buildup** is equal to the rate of **decaying** of this charge. A decrease in the hole concentration is due, first, to the injection of holes into the emitter and collector (for the general case of **double injection**) and, second, to the process of recombination in the base. The equality of the rates of positive charge buildup and decay in the base may be written in the form

$$I_b = I_{ep} + I_{cp} + \Delta Q_b / \tau_b$$

where I_{ep} and I_{cp} are hole currents in the emitter and collector respectively, and τ_b is the lifetime of carriers in the base.

Using Eq. (4.6), it is easy to express the hole currents I_{ep} and I_{cp} in terms of the excess charges ΔQ_e and ΔQ_c and the respective lifetimes τ_e and τ_c . If, for simplicity, we assume the lifetimes in all the

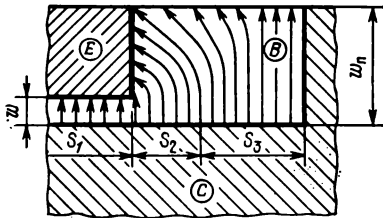


Fig. 4.10. The trajectories of electrons injected into the base in the normal region of operation

three layers of a transistor to be the same and equal to τ , then the total excess charge will be related to the base current by the elementary expression

$$\Delta Q = I_b \tau \quad (4.12)$$

With the lifetimes in the three layers being unequal, the total charge yet remains proportional to the base current, though the expression becomes too awkward. Where the charges ΔQ_e and ΔQ_c can be neglected, expression (4.12) can apply in the analysis of excess charge in the base.

To this point we have dealt with an idealized transistor structure of Fig. 4.1b. In the real structure of Fig. (4.1a), the area of the collector junction is substantially larger than the emitter junction area. Therefore, with the transistor operated in the inverted mode, the collector injects electrons not only in the active but also in the passive region of the base. The total charge stored in both portions is certainly larger than in the case of normal operation.

Since the passive part of the base is much thicker than the active, whereas the boundary concentrations of excess electrons are equal,

the carrier distribution in the passive region proves more sloping. The difference between the carrier distributions gives rise to the electron concentration gradient at the boundary between the active and passive regions of the base. The result is that a fraction of electrons injected into the passive region from area S_2 (Fig. 4.10) deflect from the straight trajectory and get into the side region of the emitter rather than reach the surface. In this case the problem dealt with in the analysis becomes *other than one-dimensional* and much more difficult. All the same, formula (4.12) remains adequate for determining the total charge of excess carriers in active and passive portions of the base.

4.4. Current Gain

In conventional transistor circuits, the output quantity (controlled variable) is either the collector or emitter current, and the input quantity (controlling variable) is either the base or emitter current. The relations between output and input currents are described by amplification factors.

4.4.1. General definitions. The relation between the collector and the emitter current may be written in the form¹

$$I_c = \alpha I_e \quad (4.13)$$

Here α is the *emitter current gain* which is one of the basic parameters of a transistor. This parameter is particularly suitable for use in cases where the emitter current can be regarded as a specified value, for example, in transistors connected in a CB circuit (see Fig. 4.3a). The value of alpha is very close to unity. In integrated circuit transistors, it commonly ranges from 0.990 to 0.995.

To establish the relation between the collector and base currents, we substitute in Eq. (4.13) the expression for $I_e = I_c + I_b$. The sought-for relation then readily reduces to the form

$$I_c = B I_b \quad (4.14)$$

Here B is the *base current gain*

$$B = \alpha / (1 - \alpha) \quad (4.15)$$

This parameter, being in widespread use in transistor engineering, is particularly applicable where the base current is a set value, first of all for transistors in the CE configuration (see Fig. 4.3b). The current gain B commonly lies in the range from 100 to 150. This factor grows as α approaches unity.

¹ If $I_e = 0$ (with the emitter circuit open), the reverse current flows in the circuit of the reverse-biased collector. For this reason, the right side of (4.13) should contain the summand I_G given by (3.24). But since it is much smaller even at high temperatures than the operating currents, I_G may be neglected.

Let us write the alpha in the form

$$\alpha = \frac{I_c}{I_e} = \frac{I_{en}}{I_e} \frac{I_c}{I_{en}}$$

Each of the two multipliers on the right of the equation has its own physical meaning and bears its own name.

The first factor

$$\gamma = I_{en}/I_e = I_{en}/(I_{en} + I_{ep}) \quad (4.16)$$

is called the *injection ratio*, or *emitter injection efficiency*, which determines the amount of the useful (electron) component in the total emitter current. Note that here we consider the *npn* transistor. In the *pnp* transistor, the useful component is the hole current. As pointed out earlier, it is only the electron component of current that is able to reach the collector and form the collector current.

The second factor

$$\kappa = I_c/I_{en} \quad (4.17)$$

is called the *transport factor*, which gives the fraction of injected carriers that have escaped the recombination on their way to the collector. It is only these carriers that make up the collector current.

So, the common-base current gain may be written in the form

$$\alpha = \gamma\kappa \quad (4.18)$$

Since the parameter α plays a dominant role in transistor operation, we shall consider its components in more detail.

4.4.2. Transport factor. To find the factor κ from (4.17), it is first necessary to calculate the current I_c . For diffusion transistors, distribution equation (4.2) can do for the purpose. For this we determine the concentration gradient at $x = w$, substitute it into (2.57a) and then multiply the result by the junction area S to find the current I_c . Next, from (4.17) we get

$$\kappa = \frac{1}{\cosh(w/L)} = \operatorname{sech} \frac{w}{L} \quad (4.19)$$

We have omitted the minus sign here because the positive direction of current I_c (see Fig. 4.3a) corresponds to the negative gradient of electron concentration. Eq. (4.19) is one of the fundamental expressions used in the theory of transistors.

Considering that $w \ll L$, it is possible to expand the right side of (4.19) into a series to within first two terms and obtain a more convenient expression

$$\kappa = \frac{1}{1 + 1/2(w/L)^2} \quad (4.20a)$$

Since the second term in the denominator is much less than unity, we can avail ourselves of one more approximation:

$$\kappa = 1 - 1/2 \left(\frac{w}{L} \right)^2 \quad (4.20b)$$

Thus if $w/L = 0.1$ or 0.2 , then $\kappa = 0.980$ to 0.995 .

Expressions (4.20) clearly show that *the transport factor comes more closely to unity as the diffusion length increases and the base width narrows*. As will be shown later, however, an increase in the diffusion length, that is, lifetime [see Eq. (2.66)] impairs the frequency characteristics of a transistor. That is why *the main trend today in transistor technology is toward a decrease in the base width*.

In drift transistors, the transport factor is derived in a similar manner and its expression has a similar structure:

$$\kappa = \frac{1}{1 + \frac{1}{2(\eta+1)} \left(\frac{w}{L} \right)^2} \quad (4.21a)$$

or

$$\kappa = 1 - \frac{1}{2(\eta+1)} \left(\frac{w}{L} \right)^2 \quad (4.21b)$$

These expressions differ from (4.20) in that they have an additional factor $(\eta + 1)^{-1}$ used earlier in (4.10b). Thus, while the diffusion transistor has $\kappa = 0.980$ to 0.995 , the drift transistor has $\kappa = 0.995$ to 0.999 , with the base width being the same and η equal to 3. From the physical viewpoint, an increase in the transport factor of drift transistors stems from the fact that carriers move faster in the accelerating field and have a less chance to recombine.

4.4.3. Injection efficiency. Divide the numerator and denominator in the right side of Eq. (4.16) by I_{en} . Further, substitute the currents I_{en} and I_{ep} from Eqs. (4.5) and (4.8), setting $x = 0$, and replace the ratio $\Delta p_e(0)/\Delta n_b(0)$ by the ratio N_b/N_e according to (3.14). The injection efficiency for a drift transistor will then take the form

$$\gamma = \left(1 + \frac{D_e}{D_b} \frac{w}{L_e} \frac{N_b}{N_e} \frac{1 - e^{-2\eta}}{2\eta} \right)^{-1} \quad (4.22a)$$

At $\eta > 1$, the exponential term may be disregarded. For diffusion transistors, assuming $\eta = 0$, we can write

$$\gamma = \left(1 + \frac{D_e}{D_b} \frac{w}{L_e} \frac{N_b}{N_e} \right)^{-1} \quad (4.22b)$$

It is apparent from the formula that the injection efficiency comes more closely to unity with a decrease in the base width and an increase in the difference between the boundary concentrations of impurities in the emitter and base layers. It is customary to *dope*

the emitter as heavily as possible, so that it generally turns to a semi-metal. The calculated values of γ can in this case be as high as 0.999 9 and even more.

Formulas (4.22) are derived on the assumption that the currents I_{en} and I_{ep} are purely of the injection origin and thus the recombination loss of carriers in the emitter junction region does not occur. In the microampere range, that is, at very small currents, such an assumption is not justifiable and the recombination in the space charge region has to be reckoned with. In this case, as seen from Fig. 3.9, the relation between the electron and hole components of the emitter current changes in favor of the hole component. In other words, the injection efficiency diminishes.

The fact that recombination becomes noticeably intensive just at low currents is due to the following. The carrier loss by recombination depends on the junction volume and thus is comparatively consistent. So, while the role of this loss seems insignificant against the background of heavy carrier flows, with a decrease in the flow of carriers this role grows in importance. A large share of recombination loss goes to the surface layer. Consequently, *the degree of surface finish has a crucial influence on the injection efficiency in the range of small currents.*

The injection efficiency, the recombination in the emitter junction being allowed for, commonly varies from 0.990 to 0.997 in the normal current range, and from 0.980 to 0.985 in the μA -range.

4.4.4. Current gain in normal and inverse region of transistor operation. If we multiply the transport factor [Eq. (4.21a)] by the injection efficiency [Eq. (4.22a)], neglecting the second-order term, and expand the result into a series to within first-order terms, we can find the common-base current gain:

$$\alpha = 1 - \frac{1}{2(\eta+1)} \left(\frac{w}{L_b} \right)^2 - \frac{D_e}{D_b} \frac{w}{L_e} \frac{N_b}{N_e} \frac{1-e^{-2\eta}}{2\eta} \quad (4.23)$$

Substituting the product of (4.21a) and (4.22a) into (4.15) and omitting the second-order term gives the common-emitter current gain. Let us write it in the form

$$\frac{1}{B} = \frac{1}{2(\eta+1)} \left(\frac{w}{L_b} \right)^2 + \frac{D_e}{D_b} \frac{w}{L_e} \frac{N_b}{N_e} \frac{1-e^{-2\eta}}{2\eta} \quad (4.24)$$

Relation (4.24) allows us to arrive at the following conclusions:

(a) the current gain of a transistor grows with decreasing base width;

(b) at a comparatively large width of the base, the transport factor plays a dominant role, while at a fairly small base width the prevailing factor is the injection efficiency;

(c) other things being the same, the current gain of a drift transistor is higher than that for a diffusion transistor.

The current gains for a transistor in the inverted mode of operation are not amenable to strict analysis because the processes of carrier motion in this case are two-dimensional (see Fig. 4.10). Many carriers injected from the collector into the passive region fail to get into the emitter, and recombine in the base layer and at the surface. For this reason, the inverse transport factor κ_1 depends to a large extent on the ratio of emitter area S_1 to collector area $S_1 + S_2 + S_3$, and can be much smaller than unity. If the collector junction is nearly symmetric, the inverse injection efficiency

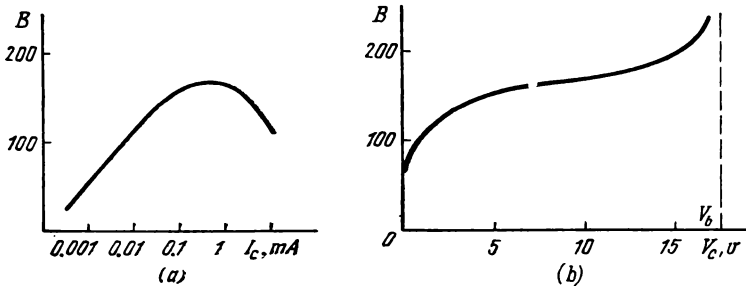


Fig. 4.11. Current gain versus collector current (a) and collector voltage (b)

γ_I will be small too. Under such conditions, the inverse current gain α_I can be as small as 0.5 and less even in diffusion transistors. In drift transistors, α_I is lower still due to the braking effect of the field.

Note, however, that inasmuch as a fair fraction of carriers injected into the passive region of the base yet reach the emitter through its side surface, it is advisable to use in the calculation of the factor κ_I a certain effective area $S_1 + S_2$ rather than a much smaller emitter bottom area S_1 alone. The estimate in this case will certainly be higher. On the other hand, one should keep in mind that the path the electrons travel until they reach the lateral face of the emitter is longer than w . This is the cause of decrease in κ_I as evident from Eq. (4.20).

Thus, the inverse parameters α_I and B_I are always lower than the normal parameters. Depending on the structure of a transistor, however, these parameters can vary in magnitude over a rather wide range. For example, B_I does not commonly exceed 0.5-1.5, whereas in special structures it runs as high as 5 to 10 and above.

4.4.5. Current gain versus temperature and operating conditions.

The current gain α and B depend on the transistor operating point (that is, on the collector current and collector voltage) and also

on temperature. The graphs of the beta values as a function of collector currents and voltages appear in Figs. 4.11 and 4.12.

The current gain fall-off in the region of small currents results from a decrease in the injection efficiency caused by recombination in the emitter junction (this fact is given due consideration in Subsec. 4.4.3). In the region of large currents (not typical of ICs), the dip of the $B = I_c$ curve is due to an increase in the base conductivity at high concentrations of excess carriers. This case is equivalent to a growth of impurity concentration in the base, which,

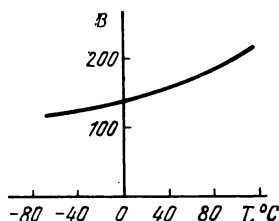


Fig. 4.12. Current gain versus temperature

according to (4.22), leads to a decrease in the injection efficiency.

In analytical form, the B - I_c relation[†] in the region of small currents can be written thus

$$B_2 = B_1 \sqrt[M]{\frac{I_{c2}}{I_{c1}}} \quad (4.25)$$

where B_1 corresponds to I_{c1} , and B_2 to I_{c2} . The index M is a specific parameter characterizing the degree of the physical and technological perfection of a transistor and its capacity to operate in the microampere range. At present it is safe to assume M approximately equal to 6, which means that B is weakly dependent on current. Until very recently, the values of M were commonly around 3 and even 2. If we set $I_{c2}/I_{c1} \approx 10^{-3}$, then $B_2 \approx 0.3B_1$ at $M = 6$; at $M = 2$, B_2 is smaller by one order of magnitude.

The $B = V_c$ relation is dependent, first, on the *Early effect* which makes itself felt at different voltages, however small, and, second, on the near-breakdown phenomena appearing in the collector junction at sufficiently high voltages.

The Early effect shows up as follows. Changes in the reverse collector voltages tend to change the collector junction width l_c see (3.9) and (3.10), which in turn causes variations in the base thickness w : if the collector junction widens, the base narrows, and vice versa (see Fig. 4.2). In the worst case, $\Delta w = -\Delta l_c$. The base-width modulation affects a number of transistor parameters, so the Early effect has often to be taken account of.

As the voltage V_c grows, the base thickness decreases on account of the Early effect and, hence, the gain B rises according to Eq. (4.24).

When V_c approaches the breakdown voltage, the collector current and thus the current gain increase still more heavily as a result of impact ionization in the collector junction (see Subsec. 3.2.7). For this range of voltages, the current gain can be written in the form

$$B = M\alpha/(1 - M\alpha) \quad (4.26)$$

where M is the impact ionization coefficient.

Given $M\alpha = 1$, when $B \rightarrow \infty$, a specific type of *breakdown* sets in which is *typical for the CE connection* of a transistor when operated at the specified base current. Equating expression (3.29) for M to $1/\alpha$, we can easily obtain the voltage at such a breakdown (see Fig. 4.11b):

$$V_B = V_M \sqrt[n]{1 - \alpha} \quad (4.27)$$

The breakdown voltage V_B is much smaller than the avalanche breakdown voltage V_M specific to CB connection (when the transistor operates at the preset emitter current). For example, if $\alpha = 0.99$ and $n = 3$, then $V_B \approx 0.2 V_M$.

A breakdown can arise not only from avalanche ionization but also as a result of base narrowing due to the rise of collector voltage (the Early effect). If the collector junction expands so that the base width becomes equal to zero, the transistor junctions will join together and the current will freely flow from the emitter to the collector, thereby causing the breakdown. This effect is known as the *reach-through (punch-through) effect*, and the voltage at which this effect takes place as the *reach-through (punch-through) voltage*. Using Eq. (3.10) for analysis, we can write the punch-through voltage in the form

$$V_w = (qN_b/2\epsilon_0\epsilon) w_0^2 \quad (4.28)$$

where N_b is the impurity concentration in the base, and w_0 is the base width at $V_c = 0$. This type of breakdown is specific to transistors with a very thin base. Thus, if $N_b = 10^{16} \text{ cm}^{-3}$ and $w_0 = 0.7 \text{ } \mu\text{m}$, then $V_w = 3.5 \text{ V}$.

The gain B-temperature relation is mainly determined by the lifetime-temperature dependence $\tau(T)$. The lifetime grows with temperature (see Fig. 2.18b) and so does the diffusion length L_b , bringing about an increase in the gain B as follows from Eq. (4.24). Besides, an increase in the lifetime retards the process of recombination in the emitter junction and thus promotes further growth of the emitter injection efficiency and the current gain B.

4.5. Static Characteristics

Proceeding from the fact that the bipolar transistor is an arrangement made up of two oppositely connected *pn* junctions, we can represent it by an equivalent circuit, or **physical model** (analog).

Fig. 4.13 shows one such model most extensively used in dc transistor circuit analysis and known as the *Ebers-Moll* model.

4.5.1. Ebers-Moll model. This model characterizes *only the active portion* of the transistor. The addition of resistors to the model, representative of the passive regions of the base and collector (see Fig. 4.1b), would make the equivalent circuit too complicated for use and less illustrative.

The Ebers-Moll model reveals well the **reversibility** of a transistor—the principal equality of its two junctions. This equality is

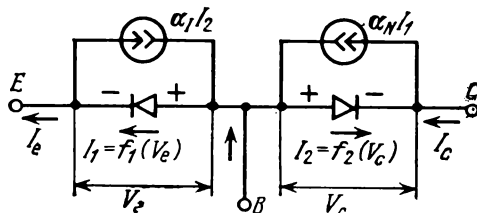


Fig. 4.13. Ebers-Moll model of a bipolar transistor

particularly evident in double injection operation of a transistor with its two junctions biased to the forward condition. In this mode of operation, each of the junctions both injects carriers into the base and collects carriers travelling from the other junction. The currents of injected carriers are designated as I_1 and I_2 , and the currents of collected carriers as $\alpha_N I_1$ and $\alpha_I I_2$, where α_N and α_I are the dc current gains in the normal and inverse regions of operation respectively. The currents $\alpha_N I_1$ and $\alpha_I I_2$ in the model under discussion are provided by current sources (generators)¹.

Write the relationships proceeding from the model of Fig. 4.13:

$$I_e = I_1 - \alpha_I I_2 \quad (4.29a)$$

$$I_c = \alpha_N I_1 - I_2 \quad (4.29b)$$

We assume that the I - V characteristic for each pn junction is described by Eq. (3.16), in which case

$$I_1 = I'_{e0} (e^{V_e / \varphi_T} - 1) \quad (4.30a)$$

$$I_2 = I'_{c0} (e^{V_c / \varphi_T} - 1) \quad (4.30b)$$

¹ The current source or current generator is a concept widely employed in circuit theory. This is a dual analog of the emf source or emf generator. An ideal emf source has zero internal impedance, while an ideal current source has infinite internal impedance: it "rigidly" sets the current in the circuit of whatever impedance.

where I'_{e0} and I'_{c0} are thermal currents at respective junctions. Each of these currents can be measured by setting the reverse voltage $|V| > 3\phi_T$ at one junction and short-circuiting the other. In practice, however, it is usual to measure thermal currents through one junction keeping the other open. The respective symbols for the currents are I_{e0} and I_{c0} .

On the strength of Eqs. (4.29) we can easily establish the relationships between thermal currents measured in the open-circuit and the short-circuit condition:

$$I'_{e0} = \frac{I_{e0}}{1 - \alpha_N \alpha_I} \quad (4.31a)$$

$$I'_{c0} = \frac{I_{c0}}{1 - \alpha_N \alpha_I} \quad (4.31b)$$

It is the quantities I_{e0} and I_{c0} which one usually refers to as thermal currents in transistor junctions.

Substituting the currents I_1 and I_2 from (4.30) into (4.29), we find analytical expressions for *static* (dc) I - V characteristics of a transistor:

$$I_e = I'_{e0} (e^{V_e/\phi_T} - 1) - \alpha_I I_{c0} (e^{V_c/\phi_T} - 1) \quad (4.32a)$$

$$I_c = \alpha_N I'_{e0} (e^{V_e/\phi_T} - 1) - I'_{c0} (e^{V_c/\phi_T} - 1) \quad (4.32b)$$

The difference between I_e and I_c readily gives the base current written as

$$I_b = (1 - \alpha_N) I'_{e0} (e^{V_e/\phi_T} - 1) + (1 - \alpha_I) I'_{c0} (e^{V_c/\phi_T} - 1) \quad (4.32c)$$

Expressions (4.32) are known as the *Ebers-Moll equations*. They represent the **mathematical model** of a bipolar transistor and are valuable in the analysis of its behavior from the active to the saturation region.

It should be pointed out that in Eqs. (4.32) **forward** voltages are considered to be positive, regardless of the fact that in *npn* transistors the polarity of forward voltages at the emitter and collector is in fact negative with respect to the base. Besides, one should keep in mind that the parameters I'_{e0} and I'_{c0} of Eqs. (4.32) are precisely *thermal currents* given by (3.17) *rather than junction reverse currents* which heavily exceed thermal currents in silicon transistors. It is only if *both* the junctions are reverse biased that formulas (4.32) become invalid. In this case the reverse currents should be estimated with due regard for the thermally generated current described by (3.26).

It can be proved that the relation

$$\alpha_N I_{e0} = \alpha_I I_{c0} \quad (4.33)$$

holds for transistors. This condition permits us to simplify some formulas obtained on the basis of Eqs. (4.32).

4.5.2. CB characteristics. As known, the independent variables for the transistor arranged in a common-base configuration are the emitter current and collector voltage. The CB characteristics, therefore, are the functions $I_c(I_e, V_c)$ and $I_e(V_e, V_c)$ represented by the families of curves. One such family representative of the function $I_c(V_c)$ with the parameter I_e (Fig. 4.14a) is known as the *set of output or collector characteristics*. The second family of curves

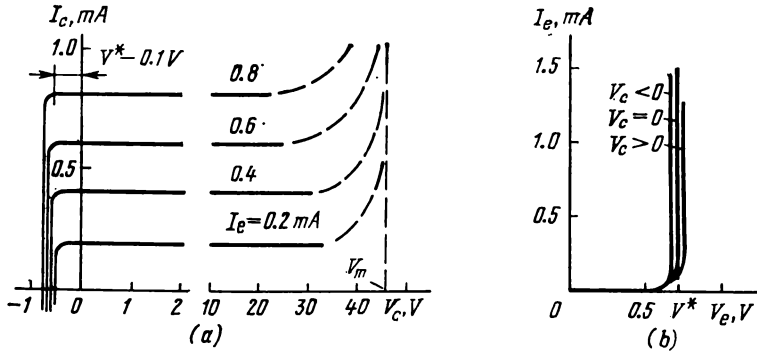


Fig. 4.14. DC output (a) and input (b) characteristics for a transistor in CB configuration

which are the plots of the function $I_e(V_e)$ with the parameters V_c (Fig. 4.14b) is the *set of input or emitter characteristics*. Both sets of curves are readily calculated from (4.32) and written in the form

$$I_c = \alpha_N I_e - I_{c0} (e^{V_c / \varphi_T} - 1) \quad (4.34)$$

$$V_e = \varphi_T \ln \left[\frac{I_e}{I_{c0}} + 1 + \alpha_N (e^{V_c / \varphi_T} - 1) \right] \quad (4.35)$$

The set of emitter characteristics of (4.35) is given as the function $V_e(I_e)$ since the *specified value* (argument) is the emitter current.

Figure 4.14a clearly illustrates two sharply different regions of transistor operation: the normal *active region* corresponding to *reverse* voltages across the collector junction (first quadrant) and the *double injection (saturation)* region corresponding to the *forward* voltage on the collector junction (second quadrant). The active mode of operation is specific to amplifying circuits, and the double injection mode to switching (pulse) circuits.

For the active mode, formulas (4.34) and (4.35) become simpler because at $|V_c| > 3\varphi_T$ the exponential terms disappear. If, in addition, we ignore the current I_{c0} and the quantity $1 - \alpha_N$, Eq. (4.34) takes the form of (4.13):

$$I_c = \alpha_N I_e \quad (4.36a)$$

and Eq. (4.35) assumes the form of (3.22):

$$V_e = \varphi_T \ln (I_e/I'_{e0}) \quad (4.36b)$$

From Eqs. (4.36) it follows that *in the active region the collector voltage does not affect either the input or the output characteristic.*

This conclusion is true for most of the practical cases. But in principle both the collector current and emitter voltage depend somewhat on the collector voltage. This means that the output characteristics of Fig. 4.14a have a finite slope determinable by the collector junction resistance given by Eq. (4.42), and the input characteristics shift somewhat with changes in collector voltage (see Fig. 4.14b). The cause of these influences is the Early effect described in Subsec. 4.4.5. The degree to which this effect causes the output characteristics to incline is discussed in Sec. 4.6. As concerns the shift of input characteristics, the Early effect shows up as follows. A change in collector voltage causes a change in the base width. Since the emitter current and thus the carrier concentration **gradient** are specified, the variation of the base width leads to a variation of the boundary carrier concentration (see Fig. 4.5). As evident from (3.12), this inevitably entails a change in the voltage across the junction.

Since Eq. (4.36b) has the same structure as general expression (3.22), it is reasonable to state once again what we have said in Subsec. 3.2.5: *in the working range of currents the voltage V_e remains almost invariable, so that it can be regarded as a parameter V^* for a silicon transistor.* The voltage V^* is equal to about 0.7 V in the normal current range (0.1 to 1 μ A) and to about 0.5 V in the micro-ampere range (1 to 10 μ A). The temperature sensitivity for emitter voltage is determined by Eq. (3.23) and ranges from $-1.5 \text{ mV } ^\circ\text{C}^{-1}$ to $-2 \text{ mV } ^\circ\text{C}^{-1}$ for silicon transistors.

A feature typical of the double injection mode is the collector current drop at an invariable emitter current. This is the result of counter-injection from the collector [see the second term on the right of Eq. (4.34)]. It should be noted that in silicon transistors a noticeable drop in I_c sets in at *sufficiently large forward voltages* V_c , rather than at $V_c = 0$. The reason is that the silicon *pn* junction (the collector junction for the case in hand) goes fully conducting only at forward voltages $V^* - 0.1 \text{ V}$, that is, at 0.4 to 0.6 V (see Subsec. 3.2.5). Till this takes place, the second term on the right of Eq. (4.34) remains negligible and the collector current stays at $\alpha_N I_e$.

4.5.3. CE characteristics. For CE transistor configuration (see Fig. 4.3b), the independent variable is the base current. Therefore, the output (collector) characteristics represent the function $I_c(I_b, V_{ce})$ and the input (base) characteristics are the function $I_b(V_b, V_{ce})$.

These characteristics, shown in Fig. 4.15, are not difficult to calculate using the Ebers-Moll equations. The main feature of the output characteristics is that they *lie entirely in the first quadrant*.

Estimate the voltage at which the collector current starts to decay. For the double injection operating condition, we can write

$$V_{ce} = V_e - V_c \quad (4.37)$$

where V_e and V_c are understood to be **forward** voltages. Formally, the edge of the active region lies at $V_c = 0$, and so the output voltage according to (4.37) is still rather high and equal to the voltage

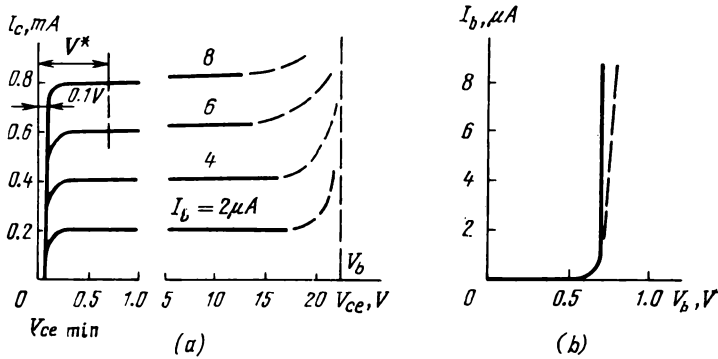


Fig. 4.15. DC output (a) and input (b) characteristics for a transistor in CE configuration

across the forward-biased emitter junction: $V_{ce} = V^* = 0.7$ V. A noticeable drop in current becomes evident only when the forward voltage V_c reaches the value $V^* - 0.1$ V (see Subsec. 3.2.5). The output voltage is then $V_{ce} = V^* - (V^* - 0.1 \text{ V}) \approx 0.1$ V (Fig. 4.15a).

A minimum value of output voltage is seen to be at zero collector current (Fig. 4.15a). To determine the value of $V_{ce \text{ min}}$, we resolve the system of equations (4.32b) and (4.32c) for voltages

$$V_e = \varphi_T \ln \left[\frac{I_b + (1 - \alpha_I) I_c}{I_{e0}} + 1 \right] \quad (4.38a)$$

$$V_c = \varphi_T \ln \left[\frac{\alpha_N I_b - (1 - \alpha_N) I_c}{I_{c0}} + 1 \right] \quad (4.38b)$$

Further, if we neglect the ones in the brackets, substitute V_e and V_c in (4.37), and use (4.33), we obtain the output voltage in the general form

$$V_{ce} = \varphi_T \ln \left[\frac{\alpha_N I_b + (1 - \alpha_I) I_c}{\alpha_I \alpha_N I_b - (1 - \alpha_N) I_c} \right] \quad (4.38c)$$

Assuming $I_c = 0$, we find the minimum output voltage:

$$V_{ce \text{ min}} = \varphi_T \ln (1/\alpha_I) \quad (4.39)$$

The voltage $V_{ce \text{ min}}$ is very small. Thus if $\alpha_I = 0.5$ (at which $B_I = 1$), $V_{ce \text{ min}} \approx 0.7 \varphi_T \approx 15 \text{ mV}$.

The slope of the I - V curve for the CE transistor configuration is much greater and the resistance that defines this slope is much smaller than is the case for the transistor in a CB connection. This is because the increment ΔV_{ce} partially drops at the emitter junction according to Eq. (4.37), bringing about an increment ΔV_e and thus ΔI_e , so that the current I_c grows **additionally**. In the near-breakdown region, the slope of the I - V curve climbs up fast. The breakdown voltage of a transistor in CE configuration is smaller than that for a common-base configuration [see Eq. (4.27)].

In conclusion let us note an important feature of the base current. As evident from Eq. (4.32c), in the normal active region, that is, at $|V_c| > 3\varphi_T$, the base current can be written as

$$I_b = I_s (e^{V_e/\varphi_T} - 1) \quad (4.40a)$$

where $I_s = (1 - \alpha_N) I'_{e0}$. Considering the recombination in the emitter junction and at the surface, however, the real current in the base takes a somewhat different form (see the dash line in Fig. 4.15b):

$$I_b = I_s (e^{V_e/m\varphi_T} - 1) \quad (4.40b)$$

where $m > 1$. The quantity m that determines the difference between the real and the ideal current is called the m -factor. This parameter is highly suitable for the estimation of *emitter junction performance* along with the level of intrinsic noise, stability, and reliability of a transistor. The m -factor is naturally related to the index M in (4.25) since M characterizes the same range of phenomena, but as applied to the injection efficiency. The relation between m and M is: $M = m/(m - 1)$. An increase in M noted above is due to a decrease in m -factor from 2 down to 1.2.

4.6. Small-Signal Circuit Models and Parameters

In a large class of the so-called linear electronic circuits, transistors operate in a specific mode such that small ac signals add to comparatively large dc components. It is just these ac signals that are of main interest in these circuits.

Write voltages and currents in the form

$$V = V^0 + \Delta V; \quad I = I^0 + \Delta I$$

where V^0 and I^0 are dc components, and ΔV and ΔI are ac components much smaller than dc components.

Direct and ac components are analyzed and calculated separately. In the analysis of direct components we have used the Ebers-Moll nonlinear physical model. For the analysis of ac components, the use of a nonlinear model is meaningless because the *relations among the small increments are determined not by the functions proper but by their derivatives*. For example, a small increment in emitter current is related to small increments in emitter and collector voltages by the expression

$$\Delta I_e = \frac{\partial I_e}{\partial V_e} \Delta V_e + \frac{\partial I_e}{\partial V_c} \Delta V_c$$

where $I_e(V_e, V_c)$ is the function given by Eq. (4.32a). For this reason the analysis of ac components relies on special **small-signal**

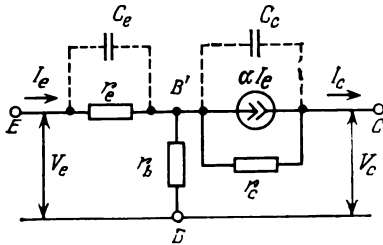


Fig. 4.16. Small-signal circuit model for a transistor in CB configuration

circuit models, or *equivalent circuits*, consisting of **linear** elements. These represent the derivatives which interrelate small increments of currents and voltages.

For a transistor having its emitter current as a parameter (the condition typical of the CB configuration), it is easy to derive the small-signal equivalent circuit from Fig. 4.13 substituting emitter and collector diodes by their incremental resistances. Since in linear electronic circuits the mode of double injection is impermissible, we can exclude the current source αI_2 from the circuit. On the other hand, the addition of the base layer resistance does not complicate the small-signal circuit, so we include the resistance r_b into the model. It would also be possible to allow for the collector series (bulk) resistance, but its role is insignificant. So the small-signal (and, in addition, **low-frequency**) equivalent circuit of the transistor with the specified emitter current takes the form like that shown in Fig. 4.16. Capacitances C_e and C_c will be covered later.

The positive direction of emitter current is chosen arbitrarily since the increment ΔI_e can be of any sign. The symbol Δ is omitted for simplicity.

Note that the current gain α (with the subscript N omitted) in the small-signal model is a **differential** (dynamic) rather than an integral (dc) factor we have used so far. The dynamic current gain α

is defined as the derivative dI_c/dI_e , whereas the dc gain α is the ratio I_c/I_e . Both factors differ in magnitude, though the difference is insignificant.

The incremental (dynamic) *emitter junction resistance* r_e is expressed in the form analogous to Eq. (3.25):

$$r_e = \varphi_T / I_e \quad (4.41)$$

where I_e is the **dc component**. At a current of 1 mA the resistance r_e is equal to 25 Ω .

The incremental (dynamic) *collector junction resistance* r_c is determined by the Early effect (see Subsec. 4.4.5). This resistance can be calculated by substituting $\alpha = \kappa$ from (4.20b) into (4.13) and differentiating the current I_c with respect to the base width w , assuming that $dw = -dl_c$ [the increment dl_c is easy to relate to dV_c with the aid of Eq. (3.10)]. The computations yield

$$r_c \approx \left(\frac{L^2}{w} \sqrt{\frac{2qN}{\epsilon_0 \epsilon}} \right) \frac{\sqrt{V_c}}{I_e} \quad (4.42)$$

where V_c is the absolute value of reverse voltage. Attention should be drawn to the fact that the *resistance* r_c , like r_e , is in inverse proportion to the dc component. Also, r_c rises a little with voltage, though this dependence is of small consequence. As an illustration, let us substitute into Eq. (4.42) the following values: $L = 10 \mu\text{m}$, $w = 1 \mu\text{m}$, $N = 10^{16} \text{ cm}^{-3}$, and $V_c = 4 \text{ V}$. We find that $r_c \approx 10^3 I_e^{-1}$; at 1 mA, $r_c = 1 \text{ M}\Omega$.

The *base resistance* r_b is, generally speaking, the sum of resistances of the active and passive regions of the base (see Fig. 4.1a). What make the calculation of these resistances a fairly laborious problem are a complex trajectory of base current, intricate geometry of the base layer and its inhomogeneity. The typical values of r_b for planar transistors can be taken to lie between 50 and 200 Ω .

If the specified input is the base current (in the CE configuration), it is more expedient to use another equivalent circuit (Fig. 4.17) with the current source in the collector circuit controlled by the base current of Eq. (4.14). Since here we deal with a small-signal circuit model, we should replace the dc current gain B by the **dynamic** (ac) current gain identified with a symbol β . The relation between small-signal parameters β and α is analogous to general expression (4.15):

$$\beta = \alpha / (1 - \alpha) \quad (4.43)$$

The factor β is greater than B in the range of low-level signal currents and smaller in the large-current range.

While replacing the current source αI_e by βI_b we should replace r_c by the quantity of a smaller value:

$$r_c^* = (1 - \alpha) r_c = r_c / (\beta + 1) \quad (4.44)$$

The quantity r_c^* owes its origin to the following considerations. To establish the parity of both equivalent circuits as fourpoles, they must have the same parameters in the open-circuit and the short-circuit condition. The open-circuit voltages in the models of

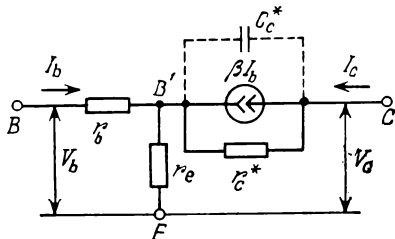


Fig. 4.17. Small-signal circuit model for a transistor in CE configuration

Figs. 4.16 and 4.17 are close to $\alpha I_e r_c$ and $\beta I_b r_c^*$ respectively. Equating these quantities and considering that in the open-circuit mode $I_e \approx I_b$, we get Eq. (4.44). As for the reason why the resistance falls off in a common-emitter transistor, we refer the reader to Subsec. 4.5.3. While r_c calculated above is equal to 1 M Ω , r_c^* is as small as 10 k Ω at $\beta = 100$.

4.7. Transient and Frequency Characteristics

The lag in the response of a transistor with fast changes of input currents is due to the path the injected carriers have to travel across the base and also due to recharge of barrier capacitances at the emitter and collector junctions. The relative role of these factors depends on the base width, transistor operating conditions, and resistance of external circuits.

Consider first the processes in the base, ignoring the effect of capacitances; their role will be discussed in Subsec. 4.7.3. Besides, we shall disregard the resistance r_c (see Fig. 14.16) since this is always the case in the analysis of transients.

4.7.1. Processes in the base of a transistor in CB connection. Let us apply a dc reverse voltage to the collector of a transistor arranged in the CB configuration (see Fig. 4.3a), keeping the emitter open. We expect a negligible thermally generated current to appear in the collector circuit. The transistor is said to stay in the *cutoff region*. At a certain moment, we allow a step of

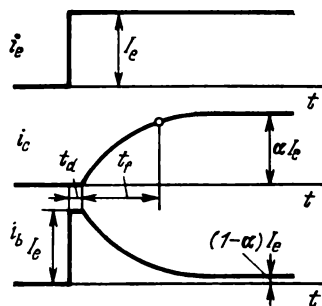


Fig. 4.18. Transients in a CB transistor

a negligible thermally generated current to appear in the collector circuit. The transistor is said to stay in the *cutoff region*. At a certain moment, we allow a step of

current, I_e , to flow through the emitter (Fig. 4.18). For simplicity, we set $\gamma = 1$, i.e. disregard the hole component of the emitter current.

Injected electrons penetrate into the base bulk gradually (see Fig. 2.27). They reach the collector only after a certain time t_d , known as the *delay time*. The collector current then starts to build up but gradually because the rate of diffusion is a **mean** value. Real diffusion rates for individual carriers differ substantially, so the carriers that have got into the base simultaneously will take

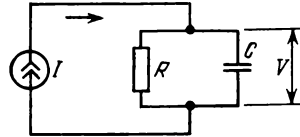


Fig. 4.19. RC circuit modeling the process of charge storage in the base

different lengths of time to reach the collector. The front of the collector current wave thus rises smoothly and has a finite duration t_f .

With a constant current applied to the emitter, the function $i_c(t)$ may conveniently be written in the form $\alpha(t) I_e$, where $\alpha(t)$ is the *step-function response of the alpha*. This characteristic is just the subject of analysis of transients in the transistor connected in a CB circuit. The parameter that determines the duration of these processes is the *time constant* τ_α being dealt with comprehensively in the text that follows.

In the interval t_d , when the collector current is still absent, the base current is equal to the emitter current I_e . Then, as the collector current grows, the base current diminishes to the steady-state value $(1 - \alpha) I_e$. So a specific initial **peak** of the base current results.

Concurrent with the increase of collector current, excess charges build up in the base. As a first approximation, which is practically justifiable, we assume the collector current and excess charges increase in an exponential manner with the time constant τ_α . The model of such a transient is the process of charging a capacitor in the simplest RC circuit on applying a step of current (Fig. 4.19). The steady-state value of charge in this circuit takes the form

$$Q = CV = C(IR) = I\tau$$

where $\tau = RC$ is the time constant. Eqs. (4.4) and (4.10b) for steady-state excess charges in the base have the same form.

Hence, a useful conclusion follows: *the time constant τ_α can be found as the quotient of the steady-state value of excess charge ΔQ_b and the specified emitter current I_e .*

Introduce the following designations for the quantities in the right side of Eqs. (4.4) and (4.10b):

diffusion time [see Eq. (3.38)]

$$t_D = w^2/2D \quad (4.45)$$

and transit time

$$t_{tr} = \frac{w^2}{2(\eta+1)D} = \frac{t_D}{\eta+1} \quad (4.46)$$

The transit time is the generalization of diffusion time for the case where the accelerating field is present in the base. If $\eta = 0$, the transit time turns equal to the diffusion time.

So, having assumed that the exponential approximation for transient characteristics is valid, we get

$$\tau_\alpha = t_{tr} \quad (4.47a)$$

or, for diffusion transistors

$$\tau_\alpha = t_D \quad (4.47b)$$

The operator transform for α is

$$\alpha(s) = \alpha/(1 + s\tau_\alpha) \quad (4.48)$$

and thus the step response (Fig. 4.20) is given by

$$\alpha(t) = \alpha(1 - e^{-t/\tau_\alpha}) \quad (4.49)$$

If we do not assume that the transients are exponential in character, the solution of the problem can take the following course.

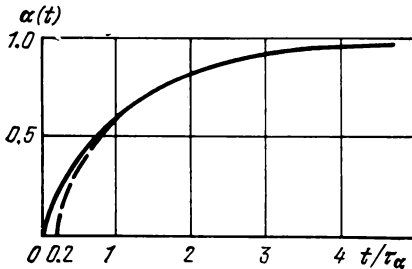


Fig. 4.20. Step response of the current gain α

As shown in Subsec. 2.8.4, the Laplace transform of the function $\Delta n(x, t)$ results from the stationary distribution $\Delta n(x)$ after replacing the diffusion length by the quantity $L(s)$ of (2.72). We apply the same approach to the problem in question.

Replacing L by $L(s)$ in (4.19) gives the strict representation $\alpha(s)$. This same representation is the transform $\alpha(s)$ since earlier we have set $\gamma = 1$. The transient characteristic corresponding to such a representation is shown in Fig. 4.20 by a dash line. Its ana-

lytical expression is too complex to be suitable for practical application. Therefore, when substituting $L(s)$ we shall make use of (4.20a) instead of (4.19). The transform $\alpha(s)$ then coincides with that of Eq. (4.48). This attests to the fact that the exponential approximation is quite appropriate for use in practice.

The shortcoming of the approximation (4.48) is that it does not take account of the delay, t_d (see Fig. 4.18). Where the delay is substantial, one should use a more accurate transform

$$\alpha(s) = \frac{\alpha e^{-st_d}}{1 + s\tau_\alpha} \quad (4.50)$$

The parameters entering this expression have the following values:

$$\tau_\alpha \approx 0.8 t_{tr} \quad (4.51a)$$

$$t_d \approx 0.2 t_{tr} \quad (4.51b)$$

The original of the transform (4.50) is the exponential function (4.49) shifted with respect to the moment $t = 0$ by the time interval t_d .

The frequency response of the current gain α can be determined by replacing in (4.48) or (4.50) the Laplace variable s by $j\omega$:

$$\dot{\alpha} = \frac{\alpha}{1 + j(\omega/\omega_\alpha)} \quad (4.52a)$$

or

$$\dot{\alpha} = \frac{\alpha e^{-j\omega t_d}}{1 + j(\omega/\omega_\alpha)} \quad (4.52b)$$

where $\omega = 1/\tau$ is the angular cutoff frequency.

It should be kept in mind that the complex quantity α can only be used in a small-signal equivalent circuit (see Fig. 4.16), that is, in the analysis of ac components. Total currents in a transistor cannot be sinusoidal in shape because of the rectifying properties of the pn junction.

The amplitude-frequency characteristics for α represented by expressions (4.52) have the same form

$$\alpha(\omega) = \frac{\alpha}{\sqrt{1 + (\omega/\omega_\alpha)^2}} \quad (4.53)$$

though the values of ω_α differ somewhat in magnitude [see Eqs. (4.47a) and (4.51a)].

The phase-frequency characteristics differ substantially. For expression (4.52a)

$$\varphi(\omega) = -\arctan(\omega/\omega_\alpha) \quad (4.54a)$$

that is, the phase shift has a limit, $\varphi(\infty) = -90^\circ$, and $\varphi(\omega_\alpha) = -45^\circ$. For expression (4.52b)

$$\varphi(\omega) = -\omega t_d - \arctan(\omega/\omega_\alpha) \quad (4.54b)$$

that is, the phase shift has no limit, $\varphi(\infty) = -\infty$, and $\varphi(\omega_\alpha) = -59^\circ$. Expression (4.54b) is much more accurate than (4.54a). The latter is responsible for a large error even at frequencies below the cutoff frequency. The frequency dependence of α is shown in Fig. 4.21.

4.7.2. Processes in the base of a transistor in CE connection. Let a step of current I_b enter into the base of a common-emitter transistor (Fig. 4.22). For the case under consideration the function $i_c(t)$ can be written as $B(t) I_b$, where $B(t)$ is the *step response* of the

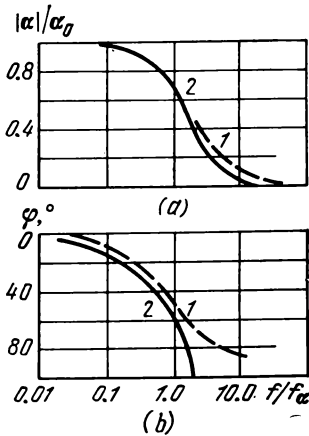


Fig. 4.21. Frequency response of the current gain α
(a) amplitude-frequency response; (b) phase-frequency response; 1—according to Eq. (4.52a); 2—according to Eq. (4.52b)

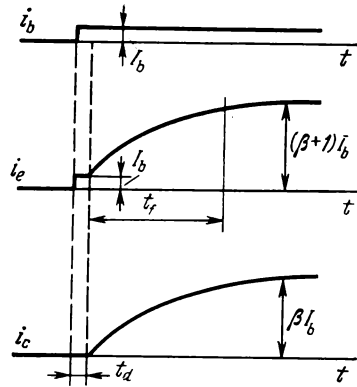


Fig. 4.22. Transients in a CE transistor

current gain B. This characteristic and its time constant τ_B are now the subjects of the analysis. For simplicity we set the injection ratio $\gamma = 1$.

The Laplace transform $B(s)$ is easy to obtain by substituting $\alpha(s)$ in (4.15). It will be readily seen that the transient here remains exponential in character, but the time constant is much larger in magnitude, namely:

$$\tau_B = \tau_\alpha / (1 - \alpha) = (B + 1) \tau_\alpha \quad (4.55)$$

To reveal the physical meaning of the quantity τ_B , let us reason out the case in the following way.

The current I_b governs the rate of growth of the positive charge in the base. So at the moment of arrival of the step input I_b , the hole concentration in the base begins to grow. Correspondingly,

the potential on the base rises, enabling the emitter junction to conduct. This triggers the injection of electrons whose charge keeps the base in a quasineutral condition. Thus the equality $I_e = I_b$ initially holds as it does in the CB configuration (see Fig. 4.18).

As the injected carriers reach the collector in a delay time t_d , the collector current appears. In the CB circuit, the buildup of collector current involves a decrease in the base current. In the CE circuit, however, the base current is constant, and so a growth in collector current (due to escape of electrons from the base) causes an additional growth in emitter current (that is an inflow of new electrons required to maintain the quasineutral state).

Such a mutual increase in I_c and I_e continues until the base stores up a sufficiently large excess charge ΔQ_b such that the rate of its recombination, $\Delta Q_b/\tau$, balances out the base current. The equilibrium condition has the form

$$\Delta Q_b \approx I_b \tau \quad (4.56)$$

Hence, being guided by the fact that $\Delta Q_b/I_b$ is the time constant τ_B of the transient (see Fig. 4.19 and its analysis), we arrive at the conclusion: *in the CE configuration, the time constant is equal to the lifetime of carriers in the base*

$$\tau_B = \tau \quad (4.57)$$

Expression (4.57) is also possible to obtain directly from Eq. (4.55) if we apply Eq. (4.21b) for α , Eq. (4.46) for τ_α , and take account of (2.66) and (4.45).

So the current gain factor in operator form for the CE configuration is written as

$$B(s) = B/(1 + s\tau_B) \quad (4.58)$$

where the time constant τ_B is tens of times as large as the quantity τ_α , and even more. It is safe to say that *the current gain B in the CE configuration grows high at a sacrifice in transient and frequency response*. The delay time t_d is insignificant as against the large time constant τ_B and thus can be neglected.

Replacing B by β and s by $j\omega$ in (4.58) gives the small-signal frequency response

$$\dot{\beta} = \frac{\beta}{1 + j(\omega/\omega_\beta)} \quad (4.59)$$

where $\omega_\beta = 1/\tau_\beta$ is the cut-off frequency (τ_β may be assumed equal to τ_B). Correspondingly, the amplitude-frequency and the phase-frequency response will be given by

$$\beta(\omega) = \frac{\beta}{\sqrt{1 + (\omega/\omega_\beta)^2}} \quad (4.60a)$$

$$\varphi(\omega) = -\arctan(\omega/\omega_\beta) \quad (4.60b)$$

Since the factor β is rather high, a transistor retains its amplifying capacity at frequencies much in excess of the frequency ω_β . At $\omega > 3\omega_\beta$, we may omit unity in (4.60a), which then reduces to

$$\beta(\omega) \approx \beta(\omega_\beta/\omega)$$

Setting $\beta(\omega) = 1$, we find the frequency at which the current gain β drops to unity and the transistor loses its ability to amplify:

$$\omega_T \approx \beta\omega_\beta = \beta/\tau \quad (4.61)$$

The frequency ω_T is called the *current-gain cutoff frequency*. Considering Eq. (4.55), we come to the conclusion that the cutoff frequency is practically equal to the frequency τ_α .

4.7.3. Effect of barrier capacitances. We shall start with the effect of emitter junction capacitance (see Fig. 4.16). Since the barrier capacitance depends on variations in the junction width, that is, on the transport of **majority** carriers, the recharging current for this capacitance has nothing to do with injection and does not enter into the composition of the collector current. Consequently, *that fraction of the emitter current which branches out into the barrier capacitance leads to a decrease in the injection efficiency.*

The distribution of emitter current between the barrier capacitance and junction depends on the relation between the resistances of these two circuits. Let us restrict ourselves to ac components. In this case we should take the quantity r_e as the junction resistance, and the impedance $1/(sC_e)$ in its operator form as the capacitive impedance. Write the fraction of current I_e that branches out into the *pn* junction:

$$I_{e\text{ pn}} = I_e \frac{1/(sC_e)}{r_e + 1/(sC_e)}$$

It is exactly this quantity that is implicit in the term emitter current I_e entering Eq. (4.16). Replacing I_e by $I_{e\text{ pn}}$ in Eq. (4.16) readily yields

$$\gamma(s) = I_{en}/I_e = \gamma/(1 + s\tau_v) \quad (4.62)$$

where

$$\tau_v = r_e C_e \quad (4.63)$$

is the time constant for the emitter junction and at the same time *the time constant of injection efficiency.*

If $r_e = 25 \Omega$ and $C_e = 1 \text{ pF}$, then $\tau_v = 0.025 \text{ ns}$; this value of τ_v is usually much smaller than t_{tr} . So it is justifiable to neglect τ_v or else to consider it as an **additional** delay. With a decrease in current, however, the resistance r_e grows and τ_v becomes comparable to t_{tr} . For the microampere range of currents, therefore, we

have ground to assume to a certain approximation that

$$\tau_\alpha \approx \tau_\gamma + t_{tr} \quad (4.64)$$

Thus *the role of the emitter barrier capacitance comes to increasing the time constant τ_α .*

We shall now concentrate on the effect of collector barrier capacitance C_c (see Fig. 4.16). If we short out the output and neglect the resistance r_c as before, the capacitance C_c will be connected in parallel with the base resistance r_b . The time constant of such a circuit is called the *base time constant*:

$$\tau_b = r_b C_c \quad (4.65)$$

Thus if $r_b = 100 \, \Omega$ and $C_c = 1 \, \text{pF}$, $\tau_b = 0.1 \, \text{ns}$. The sharing of current αI_e takes place between the **external** circuit (which includes the resistance r_b) and capacitance C_c . Consequently, at high frequencies *the external current I_c will always be lower than αI_e .* In particular, at τ_α taken equal to zero, it is just *the base time constant that sets the upper limit on the speed of transistor response.*

If the collector circuit incorporates the external resistance R_c , then for the case discussed above we should replace r_b by $r_b + R_c$. The external resistance R_c is commonly much larger than r_b , and therefore the time lag in current distribution will be determined by the time constant $C_c R_c$ rather than by τ_b . Though the quantity $C_c R_c$ is not the parameter of a **transistor** since it depends on the **external component R_c** , it is convenient to regard this quantity as a constituent part of the parameter τ_α . For this let us introduce the notion of an *equivalent time constant $\tau_{\alpha oe}$* and express it, in analogy to (4.64), as a sum

$$\tau_{\alpha oe} = \tau_\alpha + C_c R_c \quad (4.66)$$

If $C_c R_c > 3\tau_\alpha$, as is often the case, then $\tau_{\alpha oe}$ is practically independent of processes in the base.

For the CE configuration, we find the equivalent time constant τ_{oe} with the aid of Eq. (4.55), multiplying both sides of (4.66) by $B + 1$:

$$\tau_{oe} = \tau_B + C_c^* R_c \quad (4.67)$$

where

$$C_c^* = (B + 1) C_c \quad (4.68)$$

is the equivalent collector-junction capacitance (Fig. 4.17). The time constants $\tau_{\alpha oe}$ and τ_{oe} are the most universal parameters characterizing transient response of a bipolar transistor.

¹ The quantity R_c , generally speaking, also includes the intrinsic parameter of a transistor—series collector bulk resistance r_{sc} (see Fig. 4.1b). The effect of this component is described in Subsec. 7.3.3.

5.1. General

For their function unipolar transistors depend **only on one type of carrier**, the majority carriers, either electrons or holes. The processes of injection and diffusion are practically nonexistent in these transistors, or in any case they do not play a leading role. The basic type of carrier motion here is the **drift** of carriers under the effect of an electric field.

To control the current in a semiconductor at a dc field, it is necessary to change either the conductivity of the semiconductor layer or its area. Either of the approaches finds use in practice, and both

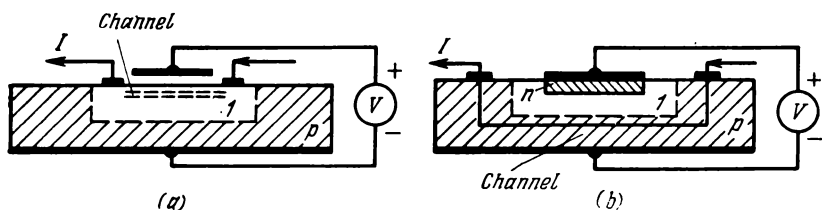


Fig. 5.1. Principle of channel action in unipolar transistors
(a) surface n channel; (b) bulk p channel; 1—depletion region

principally rely on the field effect (see Sec. 2.7). This explains why unipolar transistors are commonly called *field effect transistors* (FETs). The conducting layer through which the working current flows is termed the *channel*. Hence, one more alternative name of these transistors, *n-channel* or *p-channel* FETs.

Channels can be of the **surface** and **bulk** types. Surface channels are either enriched layers set up by donor impurities present in the insulator (see Sec. 3.4) or inversion layers resulting from the action of an external field (see Sec. 2.7). Bulk channels are the portions of a homogeneous semiconductor isolated from the surface by a depletion layer. The two types of channel and the principles of their use are illustrated in Fig. 5.1.

The surface-channel field-effect transistor (Fig. 5.1a) has a classical metal-insulator-semiconductor (MIS) structure described earlier in Sec. 2.7. Where the insulator (dielectric) is an oxide such as SiO_2 , the transistor goes under the name of *MOSFET* or just *MOST*.

The bulk-channel field-effect transistor (Fig. 5.1b) has its deple-

tion layer produced by a pn junction, for which reason it is often referred to as a pn junction FET or JFET.

Despite their difference in structure, JFETs and MOSFETs have much in common: both feature a sharply defined control circuit (with a voltage source V) clearly separated from the controlled circuit for the working current I . The control circuit does not practically consume current, since it incorporates either an insulator portion (see Fig. 5.1a) or reverse-biased pn junction (see Fig. 5.1b). The electric field set up by the control voltage has a direction normal to the flow of current. Along with the common features, each of these two types certainly shows a number of its own distinctive characteristics.

5.2. MOS Field Effect Transistors

The real structure of an n -channel MOS transistor formed from a bar-shaped slice of a p -type semiconductor is shown in Fig. 5.2. The metallic electrode producing the field effect is called the *gate* (G). The other two are the *source* (S) and *drain* (D). These two electrodes are in principle exchangeable. Of the two, the drain is the electrode to which the carriers in the channel move under the action of the applied voltage of a certain polarity. If the channel is of the n -type, the carriers are electrons and the drain is a positive electrode. The source is commonly connected to a semiconductor bar called the *substrate* (Sub).

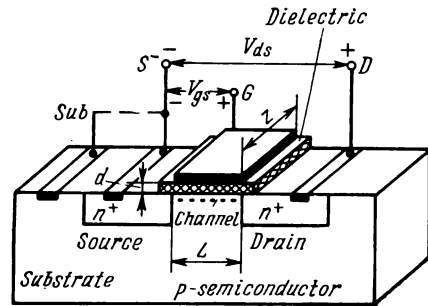


Fig. 5.2. Structure of induced n -channel MOS transistor

5.2.1. Principle of action. Ideally, when the equilibrium surface potential is equal to zero, $\varphi_{s0}=0$, an n -channel MOS transistor operates in the following manner. Assume we have the gate connected to the source, or $V_{gs} = 0$. Under these conditions the conducting channel is absent and there are two opposite-connected pn^+ junctions on the way between the source and drain. On applying the voltage V_{ds} , a negligibly small current will flow in the drain circuit.

With a negative voltage $V_{gs} < 0$ applied to the gate, the surface layer gets enriched with holes, and so the current in the drain circuit changes little. If we apply an increasing positive bias $V_{gs} > 0$ on the gate, first a depletion layer (acceptor space charge) and then an inversion layer (conducting channel) will set in (see Sec. 2.7). The drain current then assumes a definite value and only depends

on the gate voltage. The MOS transistor is now in the operating condition. Since the input current (in the gate circuit) is negligible, the device offers a considerable **power gain**, much higher than the bipolar transistor does.

The channels which are absent in the equilibrium state but appear under the effect of the externally applied voltage are known as *induced channels*. The thickness of an induced channel is practically invariable, 1 or 2 nm as noted earlier in Sec. 2.7, and thus the modulation of channel conductivity comes from changes in carrier concentration. The gate voltage at which the channel builds up is called a *threshold voltage* and denoted by V_0 . The channel length L is equal to the distance between the source and the drain region, and the channel width Z is as shown in Fig. 5.2.

The device whose structure consists of an n -type substrate with a pair of p^+ -type diffused regions (the source and drain) is a MOS transistor with an induced p channel. Typical of this device are the reverse polarities of both threshold and working voltages: $V_0 < 0$; $V_{gs} < 0$; and $V_{ds} < 0$.

Electronic circuits which use n -channel and p -channel MOS transistors in combination are called *complementary* just as are the circuits composed of nnp and pnp transistors.

It is good practice to fabricate substrates for MOS transistors from a high-conductivity material to facilitate the formation of a channel and increase the breakdown voltage of gate-source and gate-drain junctions. In principle, the mode of operation and the properties of n -channel and p -channel MOS transistors are identical, though some differences do exist. First, n -channel MOSTs show a much higher speed of response because the electrons involved in the operation of these devices have three times the mobility of holes. Second, n -channel and p -channel MOSTs differ in the structure of the surface layer in the equilibrium state, and this difference has an effect on the value of threshold voltage.

The difference in the surface layer structure is due to the different effect produced by the electrons as they leave donor impurities present in the dielectric and reach the layer (see Sec. 3.4). In the n -type substrate these electrons set up an **enriched** layer that **prevents** the formation of a p -channel; the **threshold voltage in p -channel transistors thus grows**. In the p -type substrate, the same electrons recombine with holes and produce a **depletion** layer, that is, they **encourage** the formation of an n -channel; the threshold voltage in **n -channel transistors thus decreases**.

The amount of electrons coming from the insulator can often be so high that along with the depletion layer, an inversion layer (n -channel) appears in the p -type substrate. Since it is present at zero gate voltage, this layer cannot be regarded as a channel induced by the gate fields. Hence, the ordinary meaning of threshold voltage

is of no value here. In transistors of this type, the channel available at zero gate voltage is called an *implanted* (built-in) channel; the parameter being used instead of threshold voltage is known as the *cut-off voltage*. This is the voltage which pushes off the electrons of the equilibrium inversion layer away from the surface and thus does away with the built-in channel¹.

Generally speaking, the built-in channel is not an obstacle to the application of MOS transistors. Such transistors operate **with**

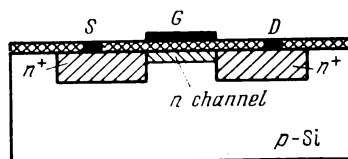


Fig. 5.3. Structure of built-in *n*-channel MOS transistor

either polarity of gate voltage: the negative voltage depletes the channel of carriers, so that the drain current decreases, while the positive voltage enriches the channel with carriers, and so the current rises. All the same, induced-channel transistors are more popular despite the fact that they are operative *only with one polarity of voltage at the gate*—such that enables the formation of the channel. Where the built-in channel is **desirable**, though this is a comparatively rare case, it can be specially obtained in the form of a thin surface layer by ion implantation (Fig. 5.3).

We shall further consider only induced *n*-channel transistors as more promising MOSFET devices inherently operating at positive working voltages, which are more convenient to deal with in analysis.

5.2.2. Threshold voltage. The gate voltage is capable of inducing a higher specific charge (the charge per unit area) in a semiconductor at a greater specific capacitance (the capacitance per unit area) between the metal and semiconductor surface. So *the gate-channel capacitance per unit area determines the controlling ability of the gate*, which feature makes this capacitance one of the important parameters of the MOS transistor. The gate-channel per-unit area capacitance is given by

$$C_0 = \epsilon_0 \epsilon_d / d \quad (5.1)$$

where d is the thickness of a dielectric (see Fig. 5.2), ϵ_0 is the electric constant, and ϵ_d is the (relative) dielectric permittivity. It is

¹ To prevent the formation of an equilibrium channel when fabricating *n*-channel MOS transistors, one has to take special measures in processing the surface of silicon and dielectric and also use bars with an increased concentration of acceptors. This makes the technology of production of *n*-channel transistors more complicated than that for *p*-channel transistors.

desirable that d of the insulator be made as thin as its breakdown voltage permits. The thickness d of a silicon oxide layer commonly varies from 0.1 to 0.15 μm . If we set d equal to 0.15 μm and ϵ_d to 3.5, then $C_0 \approx 200 \text{ pF/mm}^2$.

The threshold voltage V_0 can be divided into two components as shown on the band diagram¹ of Fig. 5.4:

$$V_0 = V_{0F} + V_{0B} \quad (5.2)$$

The component V_{0F} is responsible for *band flattening*; it reduces to zero the equilibrium surface potential φ_{s0} , that is *does away with the initial bending of bands* (see curves 1 and 2). On the diagram, the initial band bending is shown to be opposite in direction to the band bending required to build up the channel.

The component V_{0B} is responsible for *band bending* in the direction favorable for inducing the channel (curve 3): it sets up a surface potential φ_{sm} at which the level of the electrostatic potential crosses the Fermi level (see Sec. 2.7).

Thus the voltage V_{0F} determines the degree to which the semiconductor is "ready" for channel formation; if $\varphi_{s0} = 0$, then

$V_{0F} = 0$, and if the equilibrium energy bands incline downward, then $V_{0F} < 0$. As regards the voltage V_{0B} , this determines the value of threshold voltage in "ideal" conditions at which the equilibrium surface potential is equal to zero.

The voltage V_{0F} is expressed in the form

$$V_{0F} = \varphi_{MS} + Q_{0s}/C_0 \quad (5.3a)$$

where Q_{0s} is the equilibrium per-unit area surface charge which comprises the charge of surface states and the charge stemming from impurity ions in the dielectric, and φ_{MS} is the contact potential difference between metal and semiconductor. The quantity Q_{0s} is evaluated by experiment; it usually lies between $5 \times 10^{-9} \text{ C/cm}^2$ and $5 \times 10^{-8} \text{ C/cm}^2$.

The voltage V_{0B} is expressed as

$$V_{0B} = \varphi_{sm} + \frac{a}{C_0} \sqrt{\varphi_{sm}} \quad (5.3b)$$

¹ Positive values of electric potential are laid off downward.

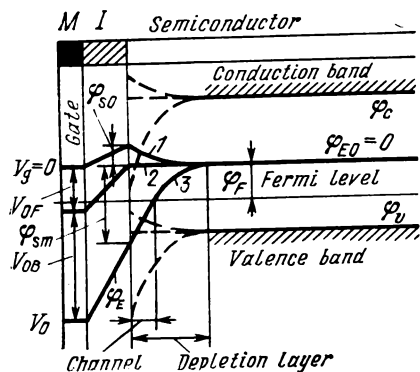


Fig. 5.4. Band diagrams for a MOS transistor at gate voltages from 0 to V_0

where

$$a = \sqrt{2q\epsilon_0\epsilon_s N} \quad (5.4)$$

is a coefficient characterizing the effect of the space charge present in the substrate. Here, ϵ_s is the permittivity of the semiconductor, and N is the impurity concentration.

It is common to set $\varphi_{sm} = 2\varphi_F$ (see p. 65), where φ_F is the difference in magnitude between the Fermi level and the electrostatic potential level in the semiconductor **bulk** (see Fig. 5.4). Thus if $N = 10^{16} \text{ cm}^{-3}$, then according to Eqs. (2.11) $\varphi_F \approx 0.3 \text{ V}$ and, hence, $\varphi_{sm} = 0.6 \text{ V}$; according to (5.4), $a \approx 5 \times 10^{-8} \text{ F V}^{1/2}/\text{cm}^2$. Assuming $C_0 = 2 \times 10^{-8} \text{ F/cm}^2$, from (5.3b) we find that $V_{0B} \approx \approx 2.6 \text{ V}$. The values of **total** threshold voltage V_0 practically lie in the range 0.5 to 3.5 V.

5.2.3. Static characteristics. Consider the effect of current on the structure of a current-conducting channel. If the drain-source voltage $V_{ds} = 0$, the semiconductor surface is **equipotential**, the field

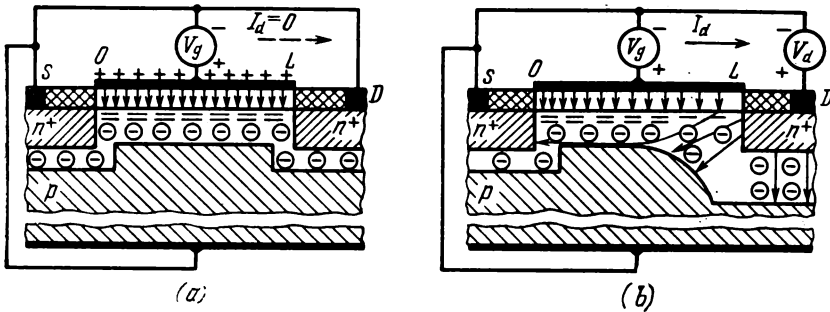


Fig. 5.5. Field and charge distribution in a MOS transistor at zero voltage (a) and low positive voltage (b) on the drain

in the dielectric is uniform, and the thickness h of the induced channel is **constant over its entire length** (Fig. 5.5a). But if $V_{ds} > 0$, the current does flow and the surface potential grows from the source toward the drain. So the potential difference between the gate and surface decreases in the direction of the drain. The field strength in the dielectric equally decreases and so does the per-unit area charge in the channel. That is why *the cross section of the channel near the point $x = L$ becomes narrower* (Fig. 5.5b).

At a certain critical voltage on the drain, called the **saturation voltage**, the potential difference between the gate and surface at the point $x = L$ becomes equal to zero. Concurrently, the field strength

in the dielectric and the specific carrier charge in the channel drop to zero at the same point (see Fig. 5.6a). This is the pinch-off condition which results in the formation of a "neck" (pinch-off region) in the channel.

The saturation voltage has the form

$$V_{d\text{ sat}} = V_{gs} - V_0 \quad (5.5)$$

At voltages $V_{ds} > V_{d\text{ sat}}$, the space charge layer, which so far has been separated from the surface by the channel, now expands to the surface over the portion ΔL , and the pinched-off end of the

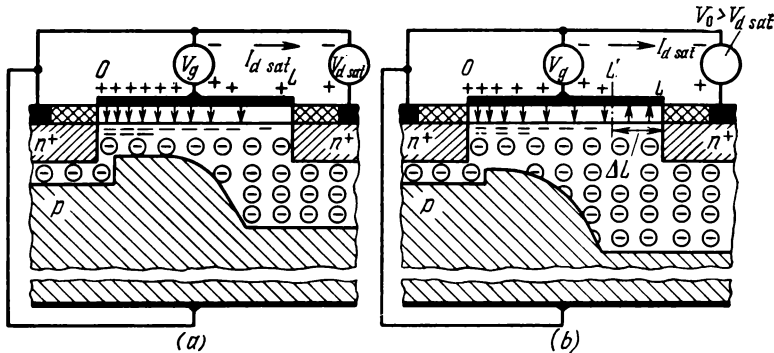


Fig. 5.6. Field and charge distribution in MOS transistor

(a) at boundary of saturation region ($V_{ds} = V_{d\text{ sat}}$); (b) in saturation region ($V_{ds} > V_{d\text{ sat}}$)

channel correspondingly shifts to the point L' as shown in Fig. 5.6b¹. Obviously, the **channel shortens** by ΔL ; the potential at point L' remains equal to $V_{d\text{ sat}}$, that is, to the same value as it was at the onset of saturation.

The value ΔL depends on the voltage difference across this portion, $V_{ds} - V_{d\text{ sat}}$. This dependence is the same in character as the dependence of the pn junction width on the reverse voltage [see Eq. (3.10)]: $\Delta L \sim \sqrt{V_{ds} - V_{d\text{ sat}}}$.

After buildup of the pinch-off region, the current in the working circuit practically ceases to depend on the drain voltage and comes to saturation (Fig. 5.7a); hence the name the saturation voltage $V_{d\text{ sat}}$.

The analysis based on the described processes gives an expression for I - V characteristics, which is too inconvenient for practical calculations (the expression contains terms raised to the two-thirds power). For this reason engineers make use of approximate expressions for I - V characteristics, of which the simplest and most popu-

¹ The processes conducive to the formation of a pinch-off region and its shift are much easier to scrutinize in junction FETs where the channel is by far thicker (see Fig. 5.13).

lar is

$$I_d = b [(V_{gs} - V_0) V_{ds} - 1/2 V_{ds}^2] \quad (5.6)$$

Here b is the *specific transconductance* of the MOS transistor, which is one of its main parameters. The expression for b is of the form

$$b = \mu C_0 \frac{Z}{L} = \frac{\epsilon_0 \epsilon \mu}{d} \frac{Z}{L} \quad (5.7)$$

where μ is the carrier mobility in the surface layer, which is commonly one-half or one-third that in the bulk, and Z is the channel

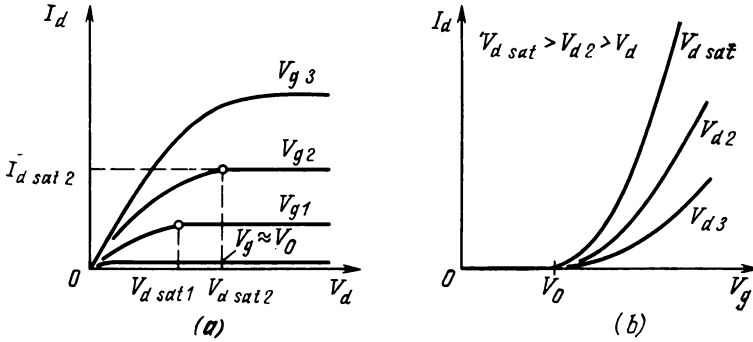


Fig. 5.7. MOS transistor static drain (a) and transfer (b) characteristics

width (see Fig. 5.2). At $\mu = 550 \text{ cm}^2/\text{V s}$, $Z/L = 10$, and $C_0 = 2 \times 10^{-8} \text{ F/cm}^2$, the typical value of specific transconductance is $b \approx 0.4 \text{ mA/V}^2$.

Expression (5.6) is valid only if $V_{ds} < V_{d,sat}$, that is, within what is called the initial, *steep* region of the I - V curves (see Fig. 5.7a). If, on the other hand, $V_{ds} > V_{d,sat}$, then the current does not vary and remains equal to the value it has at $V_{ds} = V_{d,sat}$. Substituting (5.5) into (5.6), we find the expression for the saturation region, that is, for the *flat* portions of the I - V curves:

$$I_d = 1/2 b (V_{gs} - V_0)^2 \quad (5.8)$$

This expression holds for the curve with a parameter $V_{d,sat}$ shown in Fig. 5.7b.

The *current rating* for a MOS transistor is usually considered to be the current set up at a voltage $V_{gs} = 2V_0$, that is,

$$I_{dr} = 1/2 b V_0^2 \quad (5.9)$$

As clear from the formula, *the lower the threshold voltage, the smaller the rated current*. Under nominal conditions at which $V_{gs} = 2V_0$, the saturation voltage $V_{d,sat}$ is equal to V_0 according to Eq. (5.5).

Consequently, low values of V_0 warrant both small currents and low working voltages in a transistor.

Expressions (5.6) and (5.8) enjoy extensive use for their simplicity and clarity of representation. But they lead to a substantial error in calculations if the impurity concentration in the substrate is in excess of 10^{15} cm^{-3} , which is usually the case. Therefore a more accurate approximation finds use, when necessary, in place of Eq. (5.6):

$$I_d = b [(V_{gs} - V_0) V_{ds} - 1/2 (1 + \eta) V_{ds}^2] \quad (5.10)$$

where the correction factor η is

$$\eta = \frac{1}{3} \frac{a/C_0}{\sqrt{\Phi_{sm}}} \quad (5.11)$$

Thus if $a/C_0 = 2.5 \text{ V}^{1/2}$ (the value given above) and $\Phi_{sm} = 0.6 \text{ V}$, then $\eta \approx 1.1$.

Differentiating (5.10) with respect to V_{ds} and setting $dI_d/dV_{ds} = 0$, we find the saturation voltage

$$V_{d \text{ sat}} = \frac{1}{1 + \eta} (V_{gs} - V_0) \quad (5.12)$$

This expression underestimates $V_{d \text{ sat}}$ as against (5.5). Substituting Eq. (5.12) into (5.10), we obtain a more accurate expression for the **flat** region of the I - V curves, namely, for the saturation region:

$$I_d = \frac{1}{2} \frac{b}{1 + \eta} (V_{gs} - V_0)^2 \quad (5.13)$$

The treatment so far has assumed that the source is connected to the substrate. The substrate sometimes happens to be at a **negative** potential $V_{sub \text{ s}}$ with respect to the source; an example can be ICs where the substrate is common to all transistors¹. The voltage drop in the space charge layer then increases, which makes it necessary to correct the band bending voltage given by Eq. (5.3b):

$$V_{0B} = \Phi_{sm} + \frac{a}{C_0} \sqrt{\Phi_{sm} + |V_{sub \text{ s}}|} \quad (5.14)$$

The voltage $V_{sub \text{ s}}$ will naturally enter into Eq. (5.8). Thus, the current I_d will be a **function of two voltages**, V_{gs} and $V_{sub \text{ s}}$, and so **dual control of the current** is possible.

¹ The **positive** voltage at the substrate of an n -channel transistor is impermissible, because the pn junction of the drain will operate in the forward bias state, thereby causing injection of electrons in the substrate and thus disturbing the action of the unipolar transistor.

Taking into account the substrate effect, the characteristic given by Eq. (5.13) can be expressed as

$$I_d = \frac{1}{2} \frac{b}{1+\eta} \left(V_{gs} - V_0 - \frac{2}{3} \eta |V_{sub}| \right)^2 \quad (5.15)$$

As seen, the voltage appearing between the substrate and source can be regarded as the equivalent of the threshold voltage increase.

In conclusion, consider the initial steep regions of the I - V curves which are of much importance in switching (pulse) circuits. Setting

$$V_{ds} \ll V_{gs} - V_0$$

we can neglect the quadratic term in Eq. (5.6) and obtain the linear dependence

$$I_d = b (V_{gs} - V_0) V_{ds} \quad (5.16)$$

The respective family of I - V curves is shown in Fig. 5.8. The coefficient $b (V_{gs} - V_0)$ on the right of Eq. (5.16) is known as the

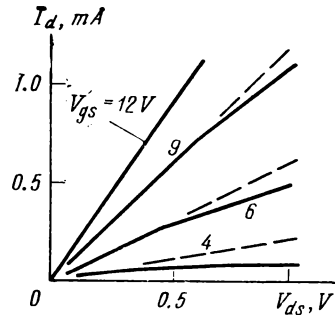


Fig. 5.8. Initial quasilinear regions of MOS transistor drain characteristics

channel conductance, and the reciprocal quantity as the *channel resistance*

$$R_0 = \frac{1}{b (V_{gs} - V_0)} \quad (5.17)$$

As clear from the formula, *the channel resistance can be controlled over a wide range by varying the gate voltage*. This feature is put to proper use in practice. If we set $V_{gs} - V_0 = 4$ V and $b = 0.1$ mA/V², then $R_0 = 2.5$ k Ω .

5.2.4. Small-signal parameters. Amplifiers are normally operated in the flat region of I - V characteristics. This region is noted for the lowest nonlinear distortion of signals and optimal small-signal parameters which are of importance for amplification.

Small-signal parameters of a MOS transistor include:

$$\text{transconductance } S = \left. \frac{dI_d}{dV_{gs}} \right|_{V_{ds}=\text{constant}}$$

$$\text{drain (internal) resistance } r_d = \left. \frac{dV_{ds}}{dI_d} \right|_{V_{gs}=\text{constant}}$$

$$\text{amplification factor}^1 k = \left. \frac{dV_{ds}}{dV_{gs}} \right|_{I_d=\text{constant}}$$

These three parameters are related by

$$k = Sr_d \quad (5.18)$$

The transconductance in the flat region is easily determined from Eq. (5.8):

$$S = b (V_{gs} - V_0) \quad (5.19a)$$

As obvious, the transconductance is proportional to the parameter b . We have assigned the term specific transconductance to the latter quantity because b is equal in magnitude to the transconductance at $V_{gs} - V_0 = 1 \text{ V}$. Using (5.19a) and (5.8) we can readily relate S to drain current:

$$S = \sqrt{2bI_d} \quad (5.19b)$$

Thus at $b = 0.1 \text{ mA/V}^2$ and $I_d = 1 \text{ mA}$, the transconductance is $S = 0.45 \text{ mA/V}$.

Formula (5.13), being more accurate than (5.19), gives a smaller value of transconductance because it contains the factor $b/(\eta + 1)$ instead of b .

The drain resistance in the flat region of the I - V curve is governed by the channel length-to-drain voltage relation (see Fig. 5.6b). The rise in voltage V_{ds} leads to an increase in the drain junction width ΔL and a respective decrease in the channel length L' . The parameter b then grows, and so does the drain current I_d .

On the whole, this phenomenon is analogous to the Early effect (see p. 132). That is why the expression for the drain resistance of a MOS transistor is similar in structure to Eq. (4.42) for the collector resistance r_c :

$$r_d = \left(L \sqrt{\frac{2qN}{\epsilon_0 \epsilon_s}} \right) \frac{\sqrt{V_d}}{I_d} \quad (5.20)$$

Assume the carrier concentration, voltage, and current to be the same as they are in the example pertaining to Eq. (4.42): $N = 10^{16} \text{ cm}^{-3}$, $V_d = 4 \text{ V}$, and $I_d = 1 \text{ mA}$. Then at $L = 10 \text{ }\mu\text{m}$, we

¹ The amplification factor of vacuum electron devices is commonly designated as μ , but since this symbol also stands for carrier mobility dealt with in the theory of MOS transistors, we replace it by k to avoid confusion.

find that $r_d = 100 \text{ k}\Omega$, which is one order of magnitude lower than r_c . Note that the r_d - I_d relation is the same as the r_c - I_c relation for bipolar transistors.

Multiplying (5.20) by (5.19b) yields the amplification factor k . This factor is independent of the channel length; its typical values range from 50 to 200 depending on the channel width Z .

It is useful to compare the parameters of bipolar and MOS transistors. Let us start with the expression for transconductance of bipolar transistors. The transconductance may readily be found from the relation

$$S = \frac{dI_c}{dV_e} = \frac{dI_c}{dI_e} \frac{dI_e}{dV_e}$$

where the first factor on the right is the current gain α , and the second factor is the reciprocal of the emitter junction resistance r_e . Thus, taking account of Eq. (4.41), we obtain

$$S = \alpha/r_e = \alpha I_e / \varphi_T$$

At $I_e = 1 \text{ mA}$, the transconductance is $S \approx 40 \text{ mA/V}$.

Multiplying S by r_c of Eq. (4.42) gives the amplification factor for a bipolar transistor:

$$k = S r_c \approx r_c / r_e$$

This coefficient does not depend on current and commonly ranges from 40 000 to 80 000.

The above examples thus reveal that MOS transistors are inferior to bipolar transistors in all the three parameters. However, one should bear in mind that the transconductance of MOS transistors rises with the channel width and, besides, is more weakly dependent on current. That is why in the range of small currents and with a large channel area, MOS transistors can be comparable to bipolar transistors in parameters.

We have mentioned above that the MOS transistor can be controlled not only by the gate voltage but also by the substrate voltage. Differentiating Eq. (5.15) with respect to $|V_{sub\ s}|$ gives the *substrate transconductance*

$$S_{sub} = -\frac{2}{3} \frac{\eta}{1+\eta} b \left(V_{gs} - V_0 - \frac{2}{3} |V_{sub\ s}| \right) \quad (5.21)$$

The minus sign says that the current I_d drops with increased voltage $|V_{sub\ s}|$. Differentiating Eq. (5.15) with respect to V_{gs} , we find the *gate transconductance*

$$S_g = \frac{b}{1+\eta} \left(V_{gs} - V_0 - \frac{2}{3} \eta |V_{sub\ s}| \right) \quad (5.22)$$

As seen, the voltage $|V_{sub\ s}|$ tends to decrease the transconductance S_g .

The relation between S_{sub} and S_g is directly dependent on the coefficient η , i.e. on the insulator thickness and impurity concentration in the substrate [see Eq. (5.11)]. Commonly, $|S_{sub}| < S_g$.

In any case, control of the drain current by the gate is **preferable** because here the input resistance conditioned by the insulator is by far higher than when effecting control by the substrate, where the input resistance is determined by the reverse current at the drain pn junction.

Let us point out in conclusion that the MOS transistor configuration considered so far, with the source common to both input and output circuits (Fig. 5.9a), is the most popular but not the solely

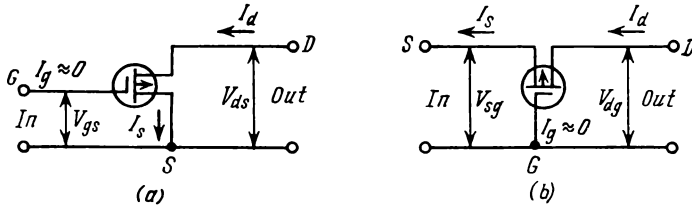


Fig. 5.9. MOS transistor in common-source configuration (a) and common-gate configuration (b)

feasible configuration. In use is sometimes the configuration with a common gate (Fig. 5.9b), which shows rather a low input resistance (close to $1/S$) and therefore finds application only in special circuits.

5.2.5. Stability of parameters. With the voltages on the gate and drain being specified, the drain current depends on temperature. This dependence shows itself as variations in the parameters b and V_0 . The function $b(T)$ is conditioned by the temperature dependence of carrier mobility, and the function $V_0(T)$ by the temperature dependence of the Fermi level [see Eq. (5.3b), where $\varphi_{sm} = 2\varphi_F$].

With an increase in temperature, both the specific transconductance and threshold voltage *decrease in magnitude*. The decreases in these parameters exert opposing effects on the drain current [see Eqs. (5.8) and (5.13)]. At a certain value of I_d the effect of $b(T)$ counterbalances the effect of $V_0(T)$. The stable current that results is called *critical*. The presence of critical current is an important feature of MOS transistors—it offers the opportunity for temperature stabilization of circuits in a simple manner, by choosing the value of working current.

Proceeding from the condition $dI_d/dT = 0$ (and allowing for derivatives $\partial b/\partial T$ and $\partial V_0/\partial T$), we can obtain the gate voltage that corresponds to the critical current

$$V_{gs\ cr} - V_0 = 0.8V \text{ to } 2.4V \quad (5.23)$$

Here the lowest value corresponds to an impurity concentration in the substrate of 10^{18} cm^{-3} , and the highest to an impurity concentration of 10^{15} cm^{-3} . The critical current is commonly one-fifth to one-tenth as high as the rated current given by Eq. (5.9).

In the current range $I_d > I_{d \text{ cr}}$ (in particular, at the rated current), the temperature coefficient of current is positive, and in the range $I_d < I_{d \text{ cr}}$ (microampere range) the TC is negative. The temperature instability is commonly defined not by the current increment but by the equivalent voltage increment ΔV_{gs} which results from the apparent relationship: $\Delta V_{gs} = \Delta I_d / S$. The temperature sensitivity for currents close to the critical value is equal to $\pm 0.5 \text{ mV } ^\circ\text{C}^{-1}$, for "supercritical" current to $+$ (8 to 10) $\text{mV } ^\circ\text{C}^{-1}$, and for "subcritical" currents to $-$ (4 to 6) $\text{mV } ^\circ\text{C}^{-1}$.

As with the drain current, the transconductance of a MOS transistor depends on temperature through the same parameters b and V_0 . Therefore, along with the concept of critical current, there is a concept of *critical transconductance*, which sets in when the effects of $b(T)$ and $V_0(T)$ on the transconductance cancel out. The critical transconductance takes place at a current below the critical value.

What affects rather heavily the parameters of a MOS transistor is that its main working region—channel—borders directly on a foreign substance, dielectric. Instability largely appears as variations in threshold voltage due mainly to changes in the equilibrium surface charge Q_{os} [see Eq. (5.3a)]. The surface charge varies, for example, because of motion of donors always present in the dielectric film. The donors can move by diffusion at high temperature or by drift in a strong gate field. Acceptor impurities that can casually get into the dielectric film or on its surface partially cancel out the donor charge, which also causes changes in the surface charge and threshold voltage.

With the current flowing in the circuit, an exchange of electrons inevitably takes place between the channel and traps lying in the dielectric film. Such an electron exchange mainly causes a random current — one of the main components of intrinsic noise in a transistor. This component relates to the category of **excess** noise attributable not to the discrete structure of the flow of carriers but to additional factors—presence of the adjacent dielectric for the case under consideration. Thus one of the limitations of MOS transistors is an increased level of intrinsic noise.

5.2.6. Transient and frequency response. A small-signal MOS transistor model in its general form is shown in Fig. 5.10a. Since the transistor is assumed to be operated in the **flat** region of the I - V characteristic, we can use the quantity r_d as the channel resistance. The elements reflecting the amplifying ability of the transistor are current sources $S_g V_{gs}$ and $S_{sub} V_{sub s}$. The resistances R_{gs} and R_{gd}

are gate insulator resistances, which are generally neglected because they range from $10^{13} \Omega$ to $10^{14} \Omega$ and over. The resistances $R_{sub s}$ and $R_{sub d}$ are the resistances of reverse biased source and drain pn

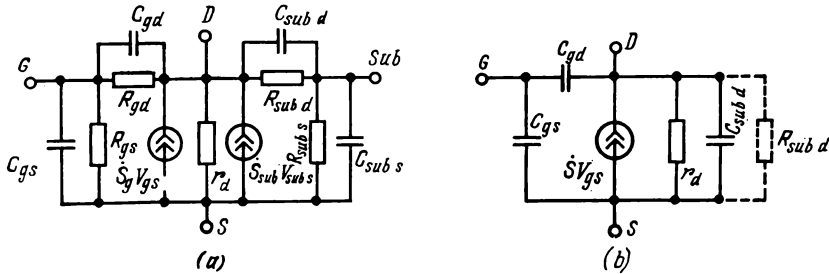


Fig. 5.10. Small-signal MOS transistor circuit models

(a) complete; (b) simplified, at $V_{sub s} = 0$

junctions; they average 10^{10} to $10^{11} \Omega$. The capacitances $C_{sub s}$ and $C_{sub d}$ are barrier (junction) capacitances of the same elements; their values depend first of all on the source and drain areas. If, for example, each of these electrodes measures 20 by 40 μm^2 , then

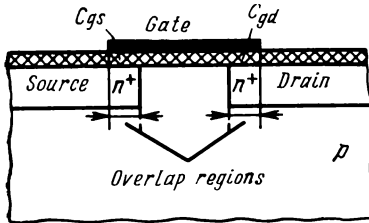


Fig. 5.11. Gate overlap and overlap capacitances

$C_{sub d} = C_{sub s} = 0.12 \text{ pF}$ at a capacitance per unit area of 150 pF (see p. 98). Finally, the capacitances C_{gs} and C_{gd} are the capacitances of the metal gate electrode with respect to the source and drain regions.

For the most widely used variant, in which the source is connected to the substrate, the current source $S_{sub}V_{sub s}$ is absent, and the resistance $R_{sub s}$ and capacitance $C_{sub s}$ are short-circuited. Besides, if the resistances R_{gs} and R_{gd} are neglected, the equivalent circuit of Fig. 5.10b results (the subscript g for S being omitted for simplicity). This circuit is the basic one for most practical calculations.

Figure 5.11 illustrates the origin of capacitances C_{gs} and C_{gd} . They stem from the so-called *overlapping of the source and drain regions by the gate* (gate overlap in short). For technological reasons it often proves impossible to place the gate electrode **exactly** between

the n^+ layers as shown in the idealized structure of Fig. 5.2. This leads to the appearance of parasitic capacitances C_{gs} and C_{gd} between the gate edges and the n^+ layers. These capacitances are normally a few times smaller than the barrier capacitances, though the role they play, in particular the capacitance C_{gd} , is significant.

The gate-to-channel capacitance C_g is not shown on Fig. 5.10 since the lag it introduces is reflected in the transconductance presented below.

The response lag of MOS transistors with fast changes of control voltage V_{gs} is attributed to two factors: recharging of gate-channel capacitance C_g and recharging of interelectrode capacitances.

The first factor can be explained in the following way. A step of voltage V_{gs} involves a change of the field in the dielectric **near the source**. So long as this change does not propagate up to the drain, the current I_d remains invariable. The propagation time depends on the speed of charging of C_g through the channel resistance.

As for the second factor, the explanation is the following. Even if the current I_d should grow stepwise, the voltage V_d and hence the current in the **external** circuit will rise **smoothly** as a result of the recharge of interelectrode capacitances. The rate of this recharging depends on **external** resistances and thus is independent of the properties of the transistor itself. Other things being equal, the recharging rate increases with decreasing interelectrode capacitances. In this respect, the capacitance values can serve as a measure of the transistor transient and frequency response.

From the above it is clear that the relative roles of both time lag factors are in principle not identical and the character of each largely depends on the type of circuit. All the same it is obvious that the second factor (the time of charging of C_g) is the limiting one: this factor determines the **transient** response of a MOS transistor with the drain short-circuited (when the effect of interelectrode capacitance is absent).

Strictly speaking, the drain circuit represents a system with distributed parameters. In engineering practice, however, it is expedient to approximate it by a simple RC circuit comprising the gate-channel capacitance C_g and channel resistance R_0 .

The channel resistance is given by Eq. (5.17). The gate-channel capacitance is easy to determine knowing the area of the gate, ZL , and its per-unit area capacitance [see Eq. (5.1)]:

$$C_g = \frac{\epsilon_0 \epsilon_d}{d} ZL \quad (5.24)$$

The charging and discharging of the RC circuit are described by the simplest exponential function. The same function can be applied to describe the transistor transconductance since this function

characterizes changes in current I_d with the given step of voltage V_{gs} . Consequently, the transconductance in operator form may be written as

$$S(s) = \frac{S}{1 + s\tau_s} \quad (5.25)$$

where $\tau_s = C_g R_0$ is the time constant of transconductance. In its complex form, the transconductance is

$$\dot{S} = \frac{S}{1 + j\omega/\omega_s} \quad (5.26)$$

where $\omega_s = 1/\tau_s$ is the cutoff frequency of transconductance. The modulus and phase of expression (5.26) will be respectively the amplitude-frequency and phase-frequency characteristics of transconductance.

The time constant τ_s can be readily obtained by multiplying the capacitance of (5.24) and channel resistance of (5.17). Considering Eq. (5.7), we get

$$\tau_s = \frac{L^2}{\mu (V_{gs} - V_0)} \quad (5.27)$$

Thus if $L = 10 \mu\text{m}$, $\mu = 500 \text{ cm}^2/\text{V s}$, and $V_{gs} - V_0 = 4 \text{ V}$, then $\tau_s = 0.5 \text{ ns}$. Hence $f_s = (1/2\pi) \omega_s \approx 300 \text{ MHz}$.

From expression (5.27) it is obvious that n channels are more preferable than p channels since the former have higher mobility μ . Also, the expression reveals the dominant role of channel length. At the present state of the art, it has become possible to fabricate MOS transistors with a channel length shorter than $1 \mu\text{m}$, thus reducing τ_s below 0.01 ns and pushing f_s above 15 GHz . *These values of parameters often allow us to neglect the transconductance lag and consider the transient response of a MOS transistor to be determined only by the interelectrode and parasitic capacitances.*

5.3. Junction Field Effect Transistors

The idealized structure of a modern JFET is given in Fig. 5.12. Here the metal contact together with the p^+ layer plays the role of a gate, the latter being separated from the n -semiconductor by the reverse-biased pn -junction depletion region rather than by the insulator as is the case with MOS transistors.

Generally speaking, the p^+ layer is dispensable in this structure: the depletion region can also exist if metal is in immediate contact with semiconductor (see Sec. 3.3). The transistors of such a structure are known as *Schottky-barrier FETs*. The basic properties of both varieties are the same, so we shall consider only the transistors with a pn junction, the analysis of which is more illustrative.

It will be explained below that for normal operation of a field effect transistor, the thickness of the working layer under the gate (denoted a in Fig. 5.12) must be not over a few micrometers. Semiconductor crystals of such a thickness prove unsuitable for use due

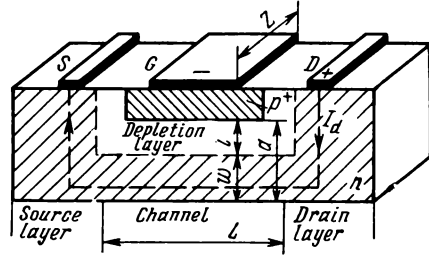


Fig. 5.12. Structure of a junction FET

to brittleness. For this reason the structure of Fig. 5.12 should be understood to be a thin n layer located on a thicker "bearing" wafer, not shown on the figure. The techniques for producing thin layers are discussed in Sec. 6.3 (see Fig. 6.3c).

5.3.1. Principle of action. The pn junction operates under reverse bias. The depletion layer depth l varies in accordance with the law described by general expression (3.9). The higher the reverse gate voltage, the deeper the depletion region and thus the smaller the channel thickness w . So, by varying the reverse gate voltage, it is possible to vary the cross section of the channel and, hence, its resistance. With the voltage present at the drain, the channel current will change, that is, the output current of the transistor.

Power gain is due to a low value of input current. In JFETs, the input current is the reverse current applied to the pn junction. For silicon pn junctions of small area, the reverse current is merely 10^{-11} A and below.

Consider now how the thickness and resistance of the channel vary with the control voltage on the gate at zero drain voltage. The channel thickness, as shown in Fig. 5.12, can be written in the form

$$w = a - l$$

where a is the distance from the "bottom" of the n layer to the metallurgical boundary of the junction. Ignoring the equilibrium height of the potential barrier and applying accordingly Eq. (3.10) for l yields the relation between the channel thickness and gate voltage:

$$w = a - \sqrt{\frac{2\epsilon_0 e V_{gs}}{qN}} \quad (5.28)$$

By V_{gs} here and below we shall mean the modulus of voltage on the gate.

Given the condition $w = 0$, it is easy to determine the *cut-off voltage* at which the depletion layer extends across the whole of the channel and the current in the channel stops flowing:

$$V_{g0} = \frac{qN}{2\epsilon_0\epsilon} a^2 \quad (5.29)$$

For example, at $N = 5 \times 10^{15} \text{ cm}^{-3}$ and $a = 2 \text{ } \mu\text{m}$, the cut-off voltage $V_{g0} = 12.5 \text{ V}$. This voltage will be slightly lower if we allow for the potential barrier height.

As apparent from the above, *the thickness of the working layer and impurity concentration in it must be fairly small*, otherwise the cut-off voltage may grow high so that complete control of current, starting from the zero value, will practically be impossible.

Using the quantity V_{g0} , the channel thickness may be written as

$$w = a \left(1 - \sqrt{\frac{V_{gs}}{V_{g0}}} \right) \quad (5.30)$$

This thickness remains a constant value over the channel length.

The channel resistance in this case is

$$R_0 = \frac{\rho L}{aZ} \left(1 - \sqrt{\frac{V_{gs}}{V_{g0}}} \right)^{-1} \quad (5.31)$$

where Z is the channel width shown in Fig. 5.12, ρ is the resistivity of the n layer. At $\rho = 1 \text{ } \Omega \text{ cm}$, $a = 2 \text{ } \mu\text{m}$, and $V_{gs} = 0$, the minimum value of R_0 min is $0.5 \text{ k}\Omega$. At $V_{gs}/V_{g0} = 0.5$, R_0 rises up to $1.8 \text{ k}\Omega$.

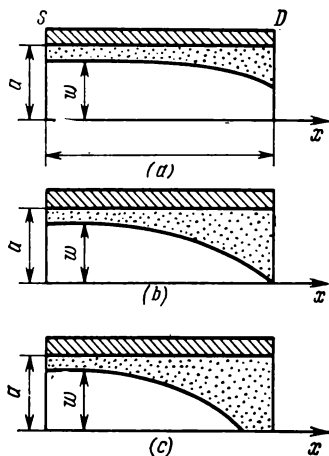


Fig. 5.13. Cross section of the channel of a JFET in unsaturated state (a), near saturation (b), and in saturated mode (c)

5.3.2. Static characteristics. With the voltage V_{ds} applied, a current starts flowing through the channel and its surface adjacent to the depletion region **ceases to be equipotential**. The voltage across the junction begins to vary along the x axis and **grow** near the drain. Hence the depletion region width, according to (3.10), grows from the source toward the drain (Fig. 5.13a).

When the potential difference $V_{ds} - V_{gs}$ (where $V_{gs} < 0$) reaches the cut-off voltage V_{g0} , the channel thickness near the drain comes to zero, resulting in the pinch-off of the channel (Fig. 5.13b). In distinction to the condition $V_{gs} = V_{g0}$, the now prevailing condition does not lead to the cut-off of current, because the pinch-off itself

is the consequence of current increase. What takes place here is the cut-off of current increments, that is, current saturation.

The formation of the pinch-off region (the "neck" in the channel) is known from the discussion of MOS transistors. When $V_{ds} - V_{gs} > V_{g0}$, the pinch-off region extends toward the source, and so the channel becomes a little shorter (Fig. 5.13c). This process is also typical of MOS transistors.

To sum up the above discussion, we write the JFET saturation condition for $V_{gs} < 0$ in the form

$$V_{d\text{ sat}} = V_{g0} - V_{gs} \quad (5.32)$$

The family of drain current-voltage characteristics (called output, or drain characteristics) appearing in Fig. 5.14a is similar to the

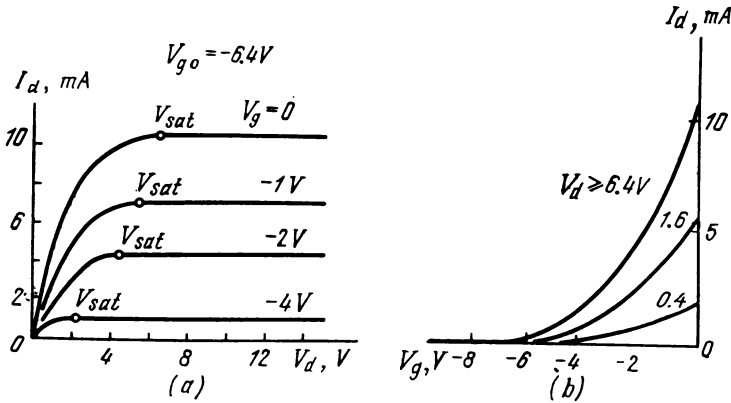


Fig. 5.14. Static drain (a) and transfer (b) characteristics of JFETs

family of I - V characteristics for MOS transistors, shown in Fig. 5.7a. However, as the gate voltage grows in magnitude, the drain current of a JFET does not rise, but falls off. We have ground to say that the JFET typically operates in the depletion region as does the built-in channel MOS transistor presented in Fig. 5.3.

The set of drain current-gate voltage characteristics (Fig. 5.14b) differs from the similar set of characteristics for MOS transistors (see Fig. 5.7b) in that the current flows at zero gate voltage. Conditionally, it is safe to say that the cut-off voltage of the JFET is equivalent to the **negative** threshold voltage of the n -channel MOSFET.

An important feature of the characteristics of Fig. 5.14b also consists in that the voltage on the gate can only be of one polarity—negative with respect to the source in the given case. Should the contrary be true, the forward voltage appearing across the pn junc-

tion would cause injection of minority carriers, thereby disturbing the normal operation of the transistor. Let us note that in built-in channel MOS transistors, which resemble JFETs, there is no limitation on the polarity of control voltage because they have the gate insulated from the channel with a dielectric material.

Analytical expressions for the I - V characteristics of a JFET are: in the steep region

$$I_d = \frac{1}{R_{0 \min}} \left[V_{ds} + \frac{2}{3} \frac{V_{gs}^{3/2} - (V_{gs} + V_{ds})^{3/2}}{V_{g0}^{1/2}} \right] \quad (5.33)$$

in the flat region

$$I_d = \frac{1}{R_{0 \min}} \left[\frac{1}{3} V_{g0} - V_{gs} \left(1 - \frac{2}{3} \sqrt{\frac{V_{gs}}{V_{g0}}} \right) \right] \quad (5.34)$$

where $R_{0 \min}$ is the channel minimum resistance at $V_{gs} = 0$ [see Eq. (5.31)]. Expression (5.34) approximates well to a quadratic equation similar in form to (5.8) for MOS transistors:

$$I_d = 1/2 b (V_{g0} - V_{gs})^2 \quad (5.35)$$

Here the coefficient b , analogous to the specific transconductance of a MOS transistor, has the form

$$b = \frac{4e_0 e \mu Z}{3aL} \quad (5.36)$$

Thus if $\mu = 1500 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$, $Z/L = 10$, and $a = 2 \mu\text{m}$. then $b = 0.12 \text{ mA/V}^2$. Note that in this example we have taken the mobility equal to the value specific to the semiconductor **bulk** because the channel of a junction FET does not extend toward the surface.

Junction FETs, like MOSFETs, exhibit critical current which is generally temperature-independent.

In junction FETs the critical current is due to the opposing effects of functions $b(T)$ and $V_{g0}(T)$. Here the function $b(T)$ is associated with the mobility-temperature dependence, as is the case with MOS transistors. The function $V_{g0}(T)$ does not follow from Eq. (5.29). But if we use a more accurate dependence (3.9) in deriving Eq. (5.29), then the latter will include the *equilibrium barrier height* of the pn junction, which is temperature dependent [see Eq. (3.4)]. It is the inclusion of this dependence that gives us the value of critical current.

Setting $dI_d/dT = 0$, we can calculate the gate voltage corresponding to the critical current

$$V_{g0} - V_{gs \text{ cr}} \approx 0.65 \text{ V} \quad (5.37)$$

The reader may compare this value with the values given by (5.23). The values of critical current usually lie in the microampere range.

5.3.3. Small-signal parameters and the equivalent circuit. Using the approximation (5.35) yields the expression for transconductance in the flat region, which is analogous to Eq. (5.19a):

$$S = b (V_{g0} - V_{gs}) \quad (5.38)$$

As for the dependence of transconductance on current, formula (5.19b) holds good.

The incremental drain resistance r_d is attributed to the same factor (channel length modulation) and has the same values as for MOS transistors [see Eq. (5.20)].

A small-signal circuit model for the junction FET appears in Fig. 5.15. The elements of this circuit are in essence the same as they are in the MOS transistor model of Fig. 5.10. Here, r_d is the incremental channel resistance in the flat region of the I - V characteristic, $\dot{S}V_{gs}$ is the current source representing the amplifying capability of the transistor, R_{gs} and R_{gd} are pn junction reverse resistances, and C_{gs} and C_{gd} are barrier capacitances at the side portions of the pn junction (see Fig. 5.12).

As with MOS transistors, the response lag in current is defined by the time constant of transconductance, τ_s . Here too, this parameter represents the product of channel resistance and gate-channel capacitance. Since the cross sections of the channel and depletion layer are not the same in different regions (see Fig. 5.13), we shall use **average** values of w and l ; namely, we set $w_{av} = l_{av} = 1/2 a$. The average capacitance and channel resistance will then be written as

$$\bar{C}_g = \frac{\epsilon_0 \epsilon (ZL)}{1/2 a} \quad (5.39a)$$

$$\bar{R}_0 = \rho \frac{L}{1/2 aZ} \quad (5.39b)$$

The average time constant of transconductance will take the form

$$\tau_s = 4\epsilon_0 \epsilon \rho L^2 / a^2 \quad (5.40)$$

This expression can be reduced to the form of (5.27) for MOS transistors by substituting a^2 from Eq. (5.29) into (5.40) and taking into account the relation $qN\mu = \sigma = 1/\rho$ [see Eq. (2.24)]. The expression for τ_s thus is

$$\tau_s = \frac{2L^2}{\mu V_{g0}} \quad (5.41)$$

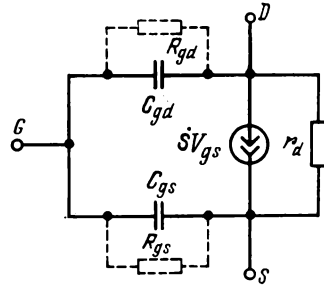


Fig. 5.15. Small-signal circuit model for a JFET

So the transient and frequency characteristics for junction FETs and MOSFETs can in principle be identical. However, transistor engineering has yet to find **practical** approaches to producing the channel length in junction FETs as short as it is in MOS transistors. This is the reason why the transient response of JFETs is much lower.

It is natural that JFETs are inferior to MOS transistors in the value of input resistance too: it results from the reverse current of the *pn* junction and does not commonly exceed $10^{11} \Omega$. As the temperature rises, the input resistance drops fast and can fall down to $10^7 \Omega$ and even below at the boundary of the operating area (125°C).

Important merits of the junction FET are a high stability of its characteristics with time and low intrinsic noise level. The reason is that the JFET has its channel isolated from the surface by the depletion layer that plays the part of a dielectric. So, at the boundary between the channel and the "dielectric" there are no crystal defects, surface conducting channels, and contaminants which are inherent in the MOS transistor and are the cause of instability and fluctuation noise. The above features also account for the fact that the carrier mobility in the channels of JFETs is equal to the bulk mobility, which is not the case with MOS transistors (see p. 157).

The only inevitable type of noise appearing in the junction FET is **thermal noise** which is specific to the channel as it is to any resistor.

Thermal noise is estimated by the Nyquist formula

$$V_{th\ n}^2 = 4kTR\Delta f$$

where Δf is a frequency band. Substituting $R_{0\ min} = 0.5\ \text{k}\Omega$ and $\Delta f = 1\ \text{Hz}$, we get $V_{th\ n} \approx 3\ \text{nV}$.

In the above analysis we have dealt only with the **active** portion of a field effect transistor—its channel. To take account of the effect of passive areas such as source and drain regions shown in Fig. 5.12, we have merely to add to the equivalent circuit resistances R_s and R_d connected in series with the source and drain. These resistances do not usually exceed 10 to 20 Ω , so their effect is of little significance as compared to that of the channel resistance.

6.1. General

Semiconductor integrated circuit technology is a logical extension of the development of transistor planar technology which embodied the prior experience gained in the production of semiconductor devices. For better understanding of the procedures of IC fabrication, therefore, we should be familiar with typical manufacturing steps of the entire technological cycle. Hybrid technology has its historical roots too. It generalized and perfected the film deposition techniques used earlier in radio engineering, machine-building industry, and optics.

6.2. Preliminary Operations

Single crystals of silicon and also other semiconductors are generally produced by the techniques of crystal growth from the melt, the most popular being the *Czochralski technique of crystal pulling* (Fig. 6.1). For a crystal to be grown, a silicon seed crystal attached to the pulling rod is lowered into contact with the melt and then slowly raised and rotated. The liquid column suspended from the seed gradually solidifies into a single crystal ingot.

The crystallographic orientation of the ingot in its cross section is defined by that of the seed. Crystals with cross sections lying in (111) or (100) planes are much more preferable than crystals of other orientations (see Sec. 2.2).

The standard diameter of crystal rods is at present 80 mm; the maximum diameter can be 120 mm and over. The length can be 1 to 1.5 m, but commonly the rods measure only fractions of this size.

Silicon ingots are first sawed into wafers, or slices, 0.4 or 0.5 mm in thickness. In the cutting operation, it is very important that the ingot should be rigidly fixed at right angles to the diamond saw blade or disk to cut out the blanks of the desired crystallographic orientation.

The surface of blanks is rather uneven: scratches, projections, and pits are far larger in size than the potential integrated elements. Before starting with basic technological steps, therefore, blanks need be repeatedly lapped and polished to produce the smooth and shiny surface. Apart from removing mechanical defects, the aim of the first stage of lapping made on special turntables is to bring the

blanks to the desired thickness, 200 to 300 μm , unattainable in sawing, and render the faces parallel to each other. The lapping agent is the suspension of micropowders chosen for each cycle of lapping in order of decreasing grain size, down to 1 or 2 μm .

The wafers lapped in this stage still have a **mechanically** disrupted surface layer, a few micrometers thick, which covers a yet thinner, **physically** disturbed layer characterized by "invisible" crystal distortions and mechanical stresses induced in the course of polishing.

Finishing polishing is aimed at removing the two disturbed layers and decreasing the surface unevenness to a level characteristic of optical systems—down to hundredths of a micrometer. This polishing can be of the **mechanical** type (polishing with yet finer-grained suspensions) and of the **chemical** type (etching of the surface layer with suitable solvents). The etch removes protrusions and cracks on the surface but has no time to dissolve the basic material, so the whole of the surface becomes more smooth.

After polishing the wafer faces become parallel to each other to within units or even fractions of a micrometer per centimeter of length.

An important process in semiconductor technology is also cleaning of the wafer surface from organic impurities, particularly from greases. The solvents used for cleaning and degreasing at increased temperature are toluene, acetone, ethanol, and others.

After etching, cleaning, and many other kinds of treatment, the wafers are rinsed in *deionized* water prepared in special units by passing the distilled water through granular resins. The chemical reactions occurring between the resins and water lead to binding of dissolved ions. The degree of deionization is estimated in terms of the resistivity of water, which usually lies in the range 10 to 20 $\text{M}\Omega\text{ cm}$ and above; the resistivity of doubly distilled water does not exceed 1 or 2 $\text{M}\Omega\text{ cm}$.

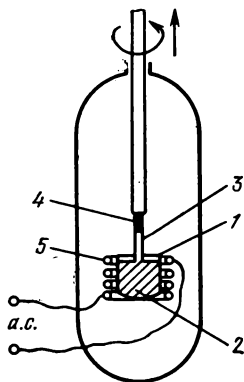


Fig. 6.1. Czochralski technique of crystal pulling
1—crucible; 2—semiconductor melt; 3—single crystal grown; 4—seed; 5—hf coil heater

6.3. Epitaxy

Epitaxy is the process of growing single crystal layers on a substrate, with the **crystallographic orientation** of the layer repeating that of the substrate material.

At present epitaxial growth techniques are generally used for depositing thin **working** layers of a homogeneous semiconductor on a comparatively thick substrate that serves as a bearing structure.

A standard process of *chloride* vapor phase growth as applied to silicon includes the following steps (Fig. 6.2). A boat-type crucible preliminarily loaded with single crystal silicon wafers is placed inside a quartz tube, and then hydrogen charged with a small amount

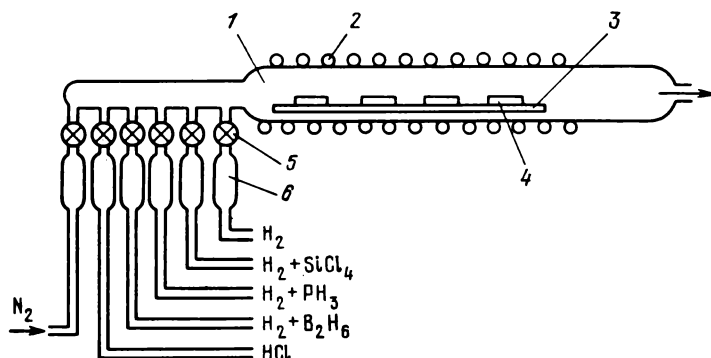
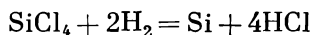


Fig. 6.2. Setup for chloride vapor phase epitaxial growth

1—quartz tube; 2—hf coil heater; 3—boat with wafers; 4—silicon wafer; 5—valve for shutting off the flow of gas; 6—gas flow rate meter

of silicon tetrachloride, SiCl₄, is passed through the tube. A high-frequency coil heater heats up the crucible to about 1200°C to trigger the chemical reaction that takes place on the surface of wafers:



A layer of pure silicon thus forms on the substrate, the vapors of HCl being carried away by the stream of hydrogen. The deposited silicon layer is a single crystal that continues the single crystal structure of the substrate. The control of temperature enables the reaction to proceed only on the wafer surface rather than in the surrounding medium.

The process occurring in the gaseous stream is called a *gas-transport reaction*, and the gas that transfers a reactant to the epitaxial growth region is known as a *carrier gas*.

If the vapors of borane, B₂H₆, or phosphorus trichloride, PH₃, are added to the vapors of silicon tetrachloride, the epitaxial layer of the intrinsic type of conductivity will turn to the layer of *p*-type or *n*-type respectively, because the acceptor atoms of boron or donor atoms of phosphorus will diffuse into the growing layer of silicon.

In the arrangement shown in Fig. 6.2, provision is made for carrying out additional operations prior to epitaxial growth. These include blow-through of the tube with nitrogen and shallow etching of the silicon surface in the vapors of HCl for its cleaning.

Thus epitaxy enables the oriented crystal growth of n - and p -type layers of any resistivity on the substrate of any type and value of conductivity (Fig. 6.3).

An epitaxial film may differ from the substrate in chemical composition. The process of growing such films is called *heteroepitaxy*

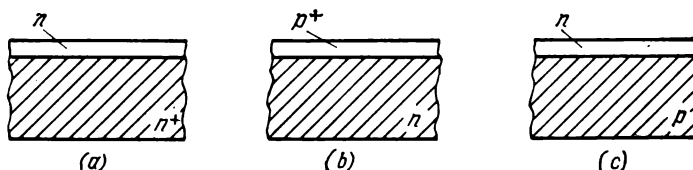


Fig. 6.3. Epitaxial structures

(a) n film on n^+ substrate; (b) p film on n substrate; (c) n film on p substrate

in contrast to *homoepitaxy* described above. Of course, the heteroepitaxial process, too, must produce the films whose crystal lattice is the same as that of the substrate. The process permits growing a silicon film on, say, a sapphire substrate.

The boundary between the epitaxial layer and substrate cannot be ideally abrupt because the impurities partially diffuse from one layer into the other in the course of epitaxy. This involves difficulties in depositing superthin (less than $1\ \mu\text{m}$) and multilayer epitaxial structures. It is the single-layer epitaxial growth that plays the leading role at present. This technique has greatly widened the scope of semiconductor technology; epitaxy can produce homogeneous layers as thin as 1 to $10\ \mu\text{m}$, unachievable so far by any other techniques.

Let us note in conclusion that along with vapor phase (gas phase) epitaxy, industry uses liquid phase epitaxy—the process of growing single crystal layers from the liquid phase, that is, from the solution containing requisite components.

6.4. Thermal Oxidation

Silicon oxidation is one of the most typical processes in modern IC technology. The process provides the film of silicon dioxide, SiO_2 , which serves a few important functions, such as:

(a) protection of the surface (through its *passivation*) and, in particular, protection of the vertical pn -junction portions getting out to the surface (Fig. 6.4a);

(b) a mask defining the windows for introduction of dopants (Fig. 6.4.b);

(c) a thin insulator under the gate of a MOS transistor (Fig. 6.4c).

The wide opportunities offered by SiO_2 are one of the reasons why silicon has become the main material for the fabrication of semiconductor ICs.

The surface of silicon is inherently coated with an oxide film resulting from natural oxidation at low temperatures. But this

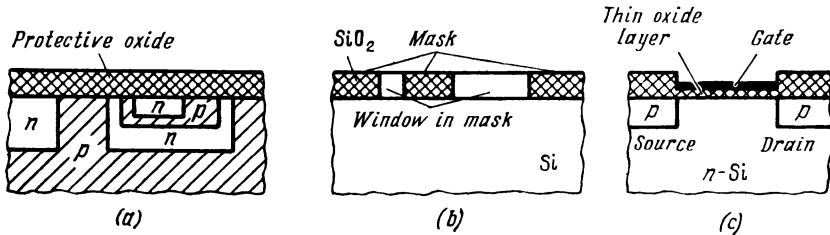


Fig. 6.4. Silicon dioxide film performing the functions of protective layer (a), mask (b) for selective doping, and thin gate oxide layer (c)

film is too thin (about 5 nm) to be able to perform any of the above functions, and therefore SiO_2 films are grown artificially at high temperatures, from 1 000 to 1 200°C.

Thermal oxidation is conducted in the atmosphere of pure oxygen (*dry oxidation*), in the mixture of oxygen and water vapors (*wet oxidation*), or just in water vapors. In wet oxidation, the oxygen-water vapor mixture is prepared by passing oxygen through a bubbler.

Oxidizing furnaces are similar to the arrangement shown in Fig. 6.2. The setup basically consists of a quartz tube to accommodate a boat with silicon wafers and a heater such as a hf coil. Dry or moistened oxygen or water vapors are forced through the tube. Oxygen reacts with silicon in the high-temperature zone to form the SiO_2 film of amorphous structure.

There are two mechanisms of oxidation. The first includes the following stages: (1) *diffusion of silicon atoms* through the natural oxide film to the surface, (2) *adsorption of oxygen molecules* by the surface from the gas phase, (3) *the oxidation proper*, or chemical reaction, which causes a film to grow **over** the initial silicon surface. The second mechanism involves (1) *adsorption of oxygen* by the surface of the natural oxide film, (2) *diffusion of oxygen* through the oxide to silicon, and (3) *the oxidation proper*. With the second mechanism, the film grows from the surface **into the bulk** of silicon. In practice, both mechanisms act in combination, but the second prevails.

The rate of oxide growth must obviously decrease with time since new oxygen atoms have to diffuse through a yet thicker layer of oxide. The semiempirical formula relating the thickness of an oxide film to the time of thermal oxidation is

$$d \approx k \sqrt{t}$$

where k is a parameter dependent on the temperature and moisture content of oxygen.

Dry oxidation proceeds tens of times slower than wet oxidation. For example, the growth of oxide film 0.5 μm thick takes about 5 h in dry oxygen at 1 000°C and merely 20 min in moistened oxygen. The time period of oxidation lengthens twofold or threefold with every decrease in temperature by 100°C.

In IC technology, it is the practice to differentiate between “thick” and “thin” oxide films. Thick oxide films with d equal to 0.7 or 0.8 μm perform the function of protection and masking, and thin films with d equal to 0.1 or 0.2 μm serve as a gate insulator in MOSFETs.

One of the important problems involved in growing the SiO_2 film is to ensure its homogeneity. Various defects may arise in the film depending on the quality of wafer surface, purity of reagents, and conditions of growing. Micropores, macropores, and even pin holes, specifically in thin films, are the widespread types of defect.

A decreased temperature of oxide growth and the use of dry oxygen give oxide films of a higher quality. For this reason, a thin gate oxide film which determines the stability of MOS transistor parameters, is grown by the dry oxidation method. In growing a thick oxide film, dry oxidation is made to alternate with wet oxidation: the first type permits avoiding defects and the second helps cut down the time needed for the entire process.

6.5. Doping

The introduction of impurities into the starting material (a wafer or epitaxial layer) by diffusion at high temperatures is still the basic method of doping of semiconductors aimed at creating diode and transistor structures. We shall focus primarily on this method. In the end of this section we shall also give due thought to ionic implantation which has gained wide acceptance in the last ten years.

6.5.1. Diffusion methods. Diffusion can be total, or overall, and selective or local. In the first case, diffusion occurs over the entire surface of the slice (Fig. 6.5a), and in the second case only in the definite portions of the slice through the windows in the mask such as the silicon oxide film (Fig. 6.5b).

Overall diffusion produces a thin diffused layer on the wafer surface that differs from the epitaxial layer by the **inhomogeneous** distribution of an impurity in depth (see the N_a - x curve in Fig. 6.5).

In **local** diffusion, the impurity penetrates not only into the wafer bulk at right angles to the wafer plane but also spreads over parallel to the wafer plane that is, **under** the mask. As a result of this **lateral diffusion**, the pn junction portion that extends outward becomes protected by the oxide (see Fig. 6.5b). The relation between the

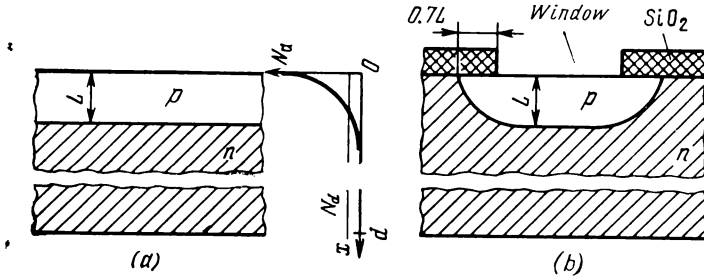


Fig. 6.5. Total (a) and selective (b) diffusion of impurity into silicon

depths of lateral and “vertical” diffusions depends on a number of factors, including the diffusion layer thickness L . The lateral diffusion depth is generally equal to $0.7 L$.

Diffusion can be performed once and repeatedly. For example, in the first stage of diffusion, it is possible to dope the starting n type slice with an acceptor impurity to produce a p layer and then, in the second stage, to drive a donor impurity into the p layer to a smaller depth and thus form a three-layer structure. So diffusion can be of the double and the triple type.

In conducting multiple diffusion, one must see that the concentration of every new impurity being introduced exceeds the preceding impurity concentration, otherwise the type of conductivity will remain the same and, hence, the pn junction will not be formed. On the other hand, the impurity concentration in silicon or any other starting material cannot be infinitely large: it has an upper limit determined by the parameter called the *solid solubility of an impurity*. The solid solubility depends on temperature. At a certain temperature, the solubility reaches its maximum, N_{\max} , and then starts to fall off. Table 6.1 gives the maximum solid solubilities of some impurities and corresponding temperatures.

In the last stage of multiple diffusion, therefore, the chosen impurity must have a maximum critical solubility. Since the range of available impurity materials is limited, it is not possible to carry out more than three diffusions in succession.

Table 6.1

Maximum Solid Solubility of Typical Impurities in Silicon

Impurity	As	P	B	Sb
N_{\max}, cm^{-3}	20×10^{20} (1 150 °C)	13×10^{20} (1 150 °C)	5×10^{20} (1 200 °C)	0.6×10^{20} (1 300 °C)

The dopants such as boron, phosphorus, and others introduced by diffusion are called *diffusants* whose **sources** are chemical compounds. These can be liquids (BBr_3 , POCl_3), solids (B_2O_3 , P_2O_5) and gases (B_2H_6 , PH_3).

As with epitaxial growth and thermal oxidation, the process of doping involves gas-transport reactions carried out in single-zone or double-zone *diffusion furnaces*.

A double-zone furnace (Fig. 6.6) consists of two high-temperature zones, one for decomposing the solid source of a diffusant and the

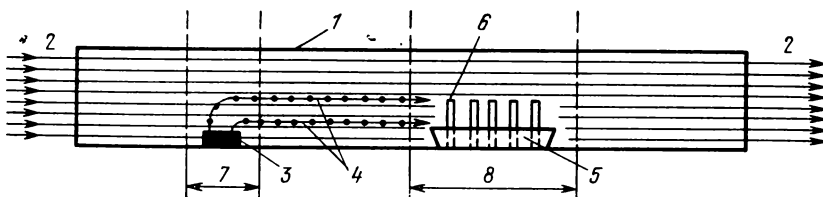


Fig. 6.6. Double-zone diffusion furnace

1—quartz tube; 2—carrier gas; 3—diffusant source; 4—diffusant source vapors; 5—crucible with wafers; 6—silicon wafer; 7—first high-temperature zone; 8—second high-temperature zone

other for performing the diffusion proper. The vapors of the diffusant source get into the stream of a neutral carrier gas, such as argon, which transfers the gaseous reactant to the second zone over the slices placed in a boat. The temperature in the second zone is higher than that in the first. The diffusant atoms penetrate the crystal lattice at the slice surface, while other components of the chemical compound are swept away from the reaction zone with the carrier gas.

Liquid and gaseous sources of a diffusant do not need high temperature for evaporation, and so they allow the use of single-zone furnaces; a diffusant source is forced into the furnace tube in the gaseous state.

Where the use is made of liquid sources of a dopant, the diffusion is performed in the oxidizing atmosphere by adding oxygen to the carrier gas. Oxygen combines chemically with the surface

atoms to form the oxide SiO_2 , which is in essence a glass. In the presence of a diffusant such as boron or phosphorus, a *borosilicate* or *phosphosilicate* glass is formed. At a temperature above $1\,000^\circ\text{C}$, these glasses are in the liquid state. They coat the silicon surface with a thin film, so that the diffusion takes place, strictly speaking, from the liquid phase. The glass solidifies to produce a sealing layer that *protects the silicon surface at the spots of diffusion*, that is, in the windows of the oxide mask. With the use of solid diffusant sources (oxides), the glass layer forms in the process of diffusion without the addition of oxygen.

6.5.2. Theory of diffusion. The basic concepts of the theory of diffusion rely on the two *laws of Fick*. Fick's first law relates the flow density of particles, J , to their concentration gradient. For the one-dimensional case,

$$J = -D (dN/dx) \quad (6.1a)$$

where D is the diffusion constant, and N is the concentration.

Fick's second law defines the rate of accumulation of particles (impurity atoms in the given case):

$$dN/\partial t = D (\partial^2 N/\partial x^2) \quad (6.1b)$$

From Eq. (6.1b) we can find the function $N(x, t)$, that is, *concentration distribution* $N(x)$ at any moment of time. For this we should set two boundary conditions.

Let the coordinate $x = 0$ correspond to the plane of the slice through which the impurity is introduced (see Fig. 6.5). The coordinate of the opposite plane will then be equal to the slice thickness d . The depth of diffused layers is practically always much less than the slice thickness (see Fig. 6.5); therefore we can assume $N(d) = 0$. From the mathematical standpoint, it is more convenient to consider the slice infinitely thick and take the following relation as the *first boundary condition*:

$$N(\infty, t) = 0 \quad (6.2)$$

The *second boundary condition* has two variants which correspond to two variants of the real manufacturing process.

1. *Unlimited impurity source.* In this case the diffusant continuously flows to the slice, so the impurity concentration in its surface layer remains constant.

The boundary condition for this variant is of the form

$$N(0, t) = N_s = \text{constant} \quad (6.3a)$$

where N_s is the *surface concentration* or, more exactly, the concentration in the surface layer. The amount of diffusant transferred to the

slice generally ensures the conditions of critical solubility, $N_s = N_{cr}$.

2. *Limited impurity source.* In this case, a certain amount of diffusant atoms are first introduced into the surface layer of the slice, and then the diffusant source is cut off to allow the *redistribution*

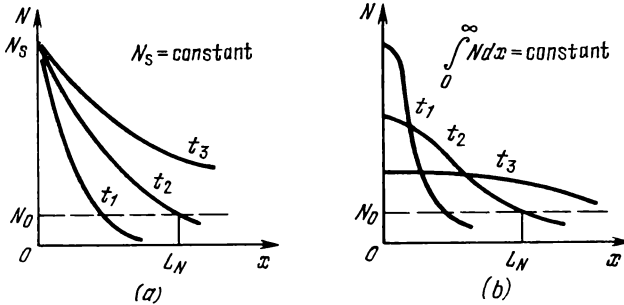


Fig. 6.7. Distribution of an impurity in the process of diffusion from the unlimited source (a) and limited source (b) as a function of time

of the *invariable quantity* of impurity atoms over the depth of the slice. The first stage represents the predeposition step, and the second the drive-in step.

For this variant we can write the condition in the form

$$\int_0^{\infty} N(x) dx = Q = \text{constant} \quad (6.3b)$$

where Q is the amount of impurity atoms *per unit area*, specified at the predeposition stage.

Solving Eq. (6.1b) for the boundary conditions (6.2) and (6.3a), we find the impurity density distribution for the unlimited source case (Fig. 6.7a):

$$N(x, t) = N_s \operatorname{erfc}(x/2\sqrt{Dt}) \quad (6.4a)$$

where $\operatorname{erfc}(z)$ is the *complementary error function* comparable to the exponential function e^{-z} (see p. 72).

The solution of (6.1b) for the conditions (6.2) and (6.3b) gives the density distribution for the **limited** impurity source (Fig. 6.7b)

$$N(x, t) = \frac{Q}{\sqrt{\pi} \sqrt{Dt}} e^{-x^2/4Dt} \quad (6.4b)$$

In the given case, the distribution follows the *Gaussian function* characteristic of which are the zero initial derivative, the inflection point, and almost an exponential tail behind this point.

By the depth of a diffused layer, or *diffusion depth*, one understands the coordinate $x = L_N$ for which the diffused impurity concentration N is equal to the initial impurity concentration N_0 (see Fig. 6.7). The quantity L_N can easily be found from Eqs. (6.4) by setting $N = N_0$.

Approximating the function (6.4a) by an exponential function, we obtain L_N for the unlimited source:

$$L_N \approx 2 \sqrt{Dt} \ln (N_s/N_0)$$

Taking the logarithms of both sides of (6.4b) gives L_N for the limited source:

$$L_N = 2 \sqrt{Dt} \ln \frac{Q}{N_0 \sqrt{\pi Dt}}$$

Both expressions have the same structure and allow us to make two important general conclusions.

1. *The diffusion time rises as the square of the diffusion depth desired, so the growth of deep diffused layers takes much time; the depth of diffused layers in ICs usually lies in the range from 1 to 4 μm .*

2. *For the given depth of a diffused layer, a change in the diffusion coefficient is equivalent to a change in the time of diffusion.*

The second conclusion deserves a more detailed consideration. Fig. 6.8 shows the temperature dependence of diffusion constants for some materials used in IC technology. It is apparent from the graphs that this dependence is exponential and thus rather strong: at $\Delta T = 100^\circ\text{C}$ the diffusion coefficient varies by a factor of 10, and at $\Delta T = \pm 1^\circ\text{C}$ it changes by $\pm 2.5\%$.

The latter variation in the diffusion coefficient would seem to be small, but a simple calculation can help understand its true significance. If $\Delta D/D = 2.5\%$, then the scatter in diffusion depth will reach $\pm 1.25\%$ or about $\pm 0.05 \mu\text{m}$ at $L_N = 4 \mu\text{m}$. So the base width $w = L_{Nb} - L_{Ne}$ can be in error to $0.1 \mu\text{m}$ or 20% at w set equal to $0.5 \mu\text{m}$. Since the coefficient β and cutoff frequency f_T are in inverse proportion to w^2 , the spread of these quantities will exceed 40% .

The above example clearly points to the need for precision control of temperature in diffusion furnaces. The permissible variation in furnace temperature lies within $\pm 0.2^\circ\text{C}$, or hundredths of a percent.

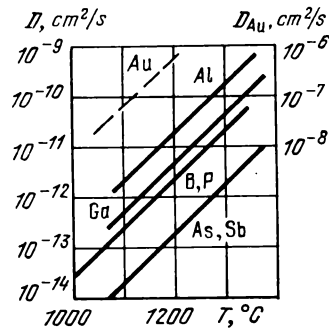


Fig. 6.8. The diffusion constants of impurities commonly introduced into silicon versus temperature (for gold, the scale is 10^{-3} times as large)

6.5.3. Ion implantation. This is the method of doping of a slice (or an epitaxial layer) by bombarding it with impurity ions accelerated to an energy enough to enable the ions to penetrate rather deep into the slice bulk.

Special installations similar to charged-particle accelerators employed in nuclear physics provide for ionization of impurity atoms, ion acceleration, and focusing of the ion beam. The dopants are the same as those used in the diffusion process.

The depth of penetration of ions depends on their energy and mass. The larger the energy of ions, the greater the thickness of the implanted layer. The increased ion energy, however, produces more *radiation defects* which impair the electrophysical properties of a crystal.

The upper limit set on the ion energy stands at 150 to 200 keV, the lower limit being set at 5 to 10 keV. In this energy range, the penetration depth reaches 0.1 to 0.4 μm , which is *much smaller than the typical depth of diffused layers*.

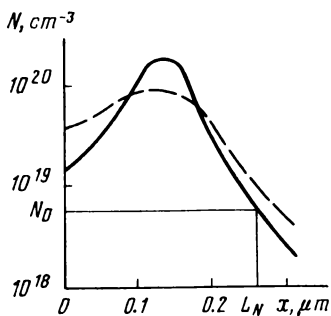


Fig. 6.9. Impurity distribution in ion implantation

The impurity concentration in the implanted layer depends on the current density in the ion beam and the process duration, or what is called the *exposure time*. The time of exposure ranges from a few seconds to 3 to 5 min and over, sometimes to 1 or 2 h, depending on the current density and the desired impurity concentration. It

stands to reason that the longer exposure of a slice to the ion beam leads to a greater number of radiation defects.

A typical impurity distribution in ion implantation is illustrated in Fig. 6.9 by a solid curve. As seen, the distribution curve shows a maximum and hence differs substantially from the curve typical of diffusion. Near the maximum, the curve is well approximated by the Gaussian function [see Eq. (6.4b)].

Since the area of the ion beam is merely 1 or 2 mm^2 , which is much smaller than the area of a slice, the ion implantation setup must have a special deflection system *to scan* the beam, that is, to deflect it smoothly or stepwise and in an adequate sequence along all the "lines" of the slice containing integrated circuits.

After completion of the doping process, the slice should be subjected to *annealing* at 500 to 800°C to ensure ordering of the silicon crystal lattice and eliminate the inevitably present radiation-induced defects, if only partially. At the annealing temperature, the diffusion processes change somewhat the profile of distribution (see the dash line of Fig. 6.9).

Ion implantation, like diffusion, can be **overall** and **local** (selective). In the latter, more typical case, the irradiation (ion bombardment) is performed through masks in which the free path of ions must be much shorter than that in silicon. The materials of masks can be silicon oxide and aluminum which find extensive uses in IC fabrication. An important merit of ion implantation is that ions, travelling along the straight line, penetrate only into the slice bulk at right angles to the surface and do not affect the regions under the mask; in other words, *the process analogous to lateral diffusion does not exist here.*

As with diffusion, multiple ion implantation for “driving” one layer into the other is in principle possible. However, it is difficult to compromise between the ion energy, exposure time, and annealing conditions required for multiple ion implantation. For this reason ion implantation enjoys popularity mainly in growing thin **single** layers.

The main advantages of ion implantation are a *low temperature* needed for the process and its *good controllability*. The first feature offers the possibility of performing ion implantation at any stage of the technological cycle, thereby dispensing with the additional diffusion of impurities into the layers prepared earlier.

6.6. Etching

Etching is usually associated with the process that uses special solutions—etches—for total or localized removal of the surface layer of a solid to the desired depth. Indeed, liquid etchants remain the basic agents for accomplishing the end. But at present new means have become available to microelectronic technology to perform the same task. In the general case, we can regard etching as a **nonmechanical** process aimed at changing the relief of the surface of solid body.

The classical process of chemical etching represents the reaction of a liquid etchant with a solid to yield a **soluble** compound mixed up with the etchant and then removed with it. Conversion of the surface layer of a solid to a solution is nothing else than elimination of this layer. But in distinction to mechanical removal of a surface layer, etching affords a far higher **precision**: dissolution occurs uniformly, the etch dissolves one monomolecular layer after the other. Choosing the right etch, its concentration, temperature and period of the etching procedure makes it possible to control rather accurately the thickness of the layer being etched. In chemical polishing of a silicon slice (see Sec. 6.2), the correctly chosen etchant can ensure an etch rate of 0.1 $\mu\text{m}/\text{min}$, thus enabling the removal of a layer merely 40 to 50 nm thick in 20 to 30 s.

For greater uniformity of etching and better expulsion of reaction products from the surface, the tray with a solution is rotated in an inclined position (*dynamic etching*), or else the solution is stirred with an ultrasonic vibrator (*ultrasonic etching*).

Of course, etching obeys the laws of physical chemistry, but in real conditions many stray factors come into play, so that the approach is to formulate etchants for each material experimentally rather than by calculations.

Local etching (through a protective mask) inherently causes the so-called lateral etching or *undercutting* (Fig. 6.10a)—the effect in a way analogous to lateral diffusion (see Fig. 6.5b). What underlies

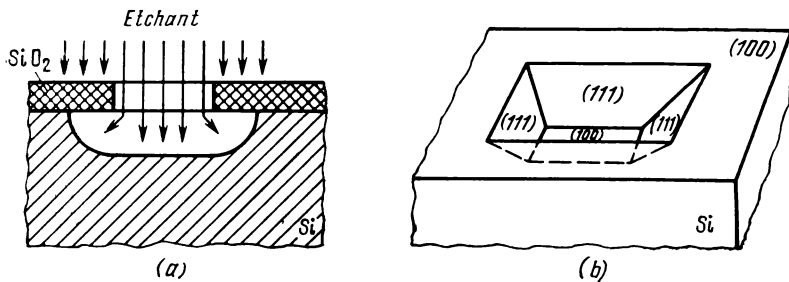


Fig. 6.10. Selective silicon etching
(a) isotropic; (b) anisotropic

this effect is that etching proceeds not only in the direction of the slice bulk but also in lateral directions, **under** the mask. As a result, the walls of the etched pit become **slightly sloping**, and the pit area a little larger than the area of the window in the mask.

Electrolytic etching is the process in which the chemical reaction of a liquid with a solid and the formation of a soluble compound occur under the effect of an electric current passed through the liquid. The solid here plays the part of one of the electrodes—anode—and thus must have a sufficient conductance, which of course places a limit on the list of materials treatable by this process. Electrolytic etching offers the advantage of controlling the etch rate by changing the current in the circuit or discontinuing the process by cutting off the current.

Ionic etching, which is a specific process applied in microelectronics, obviates the need for liquids. The etching agent here is a glow discharge produced in a vacuum just near the silicon slice being treated. The glow discharge space is filled with a *quasineutral electron plasma*. On applying to the slice a sufficiently high negative potential (with respect to the plasma), positive ions of the plasma begin to knock out atoms from the surface, layer by layer, and thus *etch*

the slice¹. In a similar manner the surface of a slice can be **cleared** of contaminants, the process being known as *ionic cleaning*. The construction of ion-plasma chambers is described in Sec. 6.9.

Ionic etching can be total and local (selective), as is chemical etching. A definite advantage of local ion etching is the absence of lateral etching under the mask: the walls of the etched pit are practically vertical and the pit area is equal to the area of the mask window.

The general advantage of ionic etching lies in its universality since the process eliminates the need for a painstaking selection of etchants for each material, the general disadvantage being that the process requires costly arrangements and takes much time to produce the desired vacuum in chambers.

Last years have seen the development and wide practical application of what is called *anisotropic etching*. The technique takes advantage of the fact that the rate of the chemical reaction underlying classical etching depends on the crystallographic orientation. The lowest rate is specific to the [111] direction, in which the density of atoms per unit area is maximum (see Fig. 2.2), and the highest rate to the [100] direction, in which the atom density is at a minimum. Special **anisotropic** etchants can dissolve a material at a different rate in different directions, so that the side walls of the pit assume a definite relief, a *cut*. Fig. 6.10*b* illustrates an example of the cut obtained in etching in the (100) plane. As seen, etching proceeds **parallel** to the (111) planes, since in the [111] orientation perpendicular to the (100) plane the etch is by far lower than in the other directions.

The angles at which the side walls of pits are etched off are strictly definite and amenable to calculation. For example, the angle between the faces (100) and (111) shown in Fig. 6.10*b* is equal to $61^{\circ}45'$. The anisotropic etching technique combined with the masking technique enables the IC design engineer to lay out the contours of pits in depth as well as along the plane.

The fact that the (111) plane is impermeable, as it were, to the etchant offers one more advantage of anisotropic etching: if the edges of windows in the mask are oriented along the (100) axes, the lateral etching effect intrinsic in isotropic etching as shown in Fig. 6.10*a* is absent. In anisotropic etching, therefore, the outer dimensions of pits may practically coincide with the dimensions of windows in the mask.

¹ The voltage in ionic etching is merely 2 or 3 keV, that is, much lower than the accelerating voltages applied in ion implantation, and so ions do not drive deep into the slice.

6.7. Masking

Masks hold an important place in the technology of semiconductor devices. They serve to ensure the local character of deposition, doping, etching, and, in some cases, epitaxial growth. Every mask has a preliminarily designed combination of openings or windows. The preparation of these windows is the task of *lithography* or engraving. Of all the techniques used for mask fabrication *photolithography* heads the list, so we shall give it primary attention.

6.7.1. Photolithography. This technique, also called *photomasking* or *photoengraving* uses *photoresists* which are a variety of photo-emulsions applied in conventional photography. Photoresists are sensitive to ultraviolet light and hence they can be processed in a slightly darkened room.

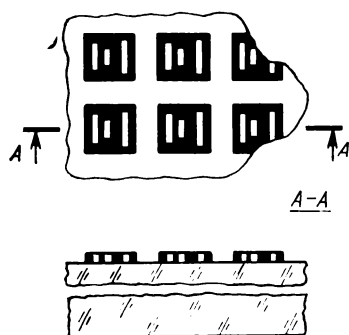


Fig. 6.11. Fragment of a photomask shown in plan and in section

Photoresists come in negative-acting and positive-acting types. The first polymerize under light and become stable to etchants (acidic or alkaline solutions); after selective exposure to light the **unexposed** portions will be soluble as is the case for a common photographic negative. On the contrary, in positive photoresists the light destroys polymer chains so the etch will dissolve the exposed portions.

The structure containing the pattern of the future oxide mask is known as a *photomask* (Fig. 6.11)

This is a thick glass plate one side of which is coated with a thin non-transparent film having the desired circuit *pattern* in the form of transparent openings. These openings or pattern elements are equal in size to the desired integrated elements, which can be as small as 20 to 50 μm or even 2 or 3 μm . Since ICs are fabricated by the **batch** technique (see Sec. 1.2) the photomask has a large number of **single-type** pattern elements arranged along the lines and in columns. Each pattern element is equal in size to the desired IC chip.

The photolitho technique for opening windows in the SiO_2 mask that covers a silicon wafer consists of a number of steps (Fig. 6.12).

A small drop of photoresist PR is placed on the oxidized surface and the wafer is rotated to spread the photoresist over its surface in an even film about 1 μm thick. The film is then left to dry hard. Next the photomask FM with its pattern facing the photoresist is placed over the wafer and exposed to the light of a quartz lamp (Fig. 6.12a). The photomask is then taken off.

If the process makes use of a positive photoresist then after its development and fixing (hardening and heat treatment) the photoresist layer will have windows in the areas which correspond to the transparent portions on the photomask¹. We thus have *transferred image* of the pattern from the photomask to the photoresist. The photoresist layer is now the mask that *tightly adheres to* the oxide layer (Fig. 6.12b).

In the next step an etchant is applied to remove the oxide layer through the windows in the photoresist mask as far as the silicon

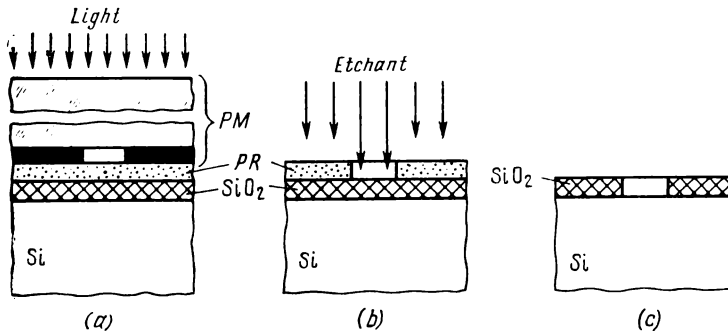


Fig. 6.12. Basic steps in the photomasking process

(a) exposure of photoresist through photomask; (b) selective etching of SiO_2 through photoresist mask; (c) oxide mask after photoresist removal

surface (which is resistant to the etchant used) and thus to open the windows in the oxide thereby transferring the pattern from the photoresist to the oxide layer. The final step involved in the photomasking process comes to etching away the remaining photoresist leaving intact the **oxide** mask with windows (Figs. 6.12c and 6.4b). The wafer is now ready for such operations as diffusion or ion implantation etching and so forth until the integrated circuits are completed.

In the technological cycles of production of diodes, transistors, and especially integrated circuits, the photomasking process need be **repeated** for the fabrication of, say, base layers, emitters, and ohmic contacts. This raises the problem of proper alignment of photomasks. Fig. 6.13 gives an example of pattern orientation.

Assume the preceding photomasking and diffusion processes have given a p layer 30 μm wide and in the next similar stage we have to drive inside the p layer a 10 μm wide n -type layer (shown by a dash line) shifted 7 μm relative to the p layer centerline. For this we should

¹ The use of a **negative** photoresist necessitates a negative photomask whose nontransparent portions must represent the desired openings in the oxide mask.

register the pattern on the second photomask with the diffused structure (p layer boundaries) to an accuracy of 1 or 2 μm .

In multiple photomasking (involving 5 to 7 photomasking operations in the IC fabrication procedure) tolerances for alignment are only fractions of a micrometer. The technique of alignment uses registration marks (positioning aids) produced in the form of crosses or squares on photomasks. These marks subsequently pass into the pattern in the oxide layer and are visible under the thin photoresist film. Locating the next photomask on the slice the operator

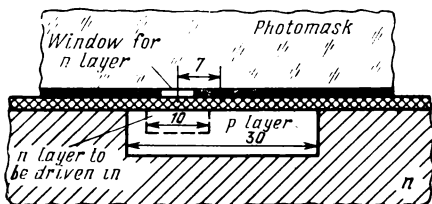


Fig. 6.13. Photomask alignment on the surface of an IC substrate

brings the mark on the photomask in register with the mark in the oxide layer in the most accurate way possible, using a microscope for the purpose.

The described photolitho technique is typical for the manufacture of **oxide masks** on silicon slices to make them ready for the subsequent selective diffusion. The photoresist mask (see Fig. 6.12b) is an **intermediate, auxiliary masking medium** since it does not withstand high temperatures common to diffusion. But in some cases where the process occurs at low temperatures the photoresist can be a **basic, working mask**. An example can be the process of producing a metallization pattern for semiconductor ICs (see Sec. 6.9).

6.7.2. Photomasks. The first stage in the process of manufacture of photomasks involves preparing an *artwork* from the completed layout design. The artwork is a drawing of one of the elements of the photomask¹ made at a scale of 100 : 1 to 1 000 : 1. So, a $10 \times 20 \mu\text{m}$ rectangle on the photomask corresponds to an artwork portion measuring $2 \times 4 \text{ mm}$ or $5 \times 10 \text{ mm}$, depending on the chosen scale. These artwork elements can be cut to a high precision, with the edges of cuts accurate to within $\pm 25 \mu\text{m}$, or a few percent. The artwork for a $1.5 \times 1.5 \text{ mm}$ chip can measure $50 \times 50 \text{ cm}$ and over.

Artworks are prepared on *coordinatographs*. Fig. 6.14 illustrates one such tool in schematic form for generating artwork. It consists

¹ The photomask element is a portion corresponding to one chip. In the batch technique of IC fabrication, the photomask contains hundreds of pattern elements to produce hundreds of integrated elements on a single slice (see Fig. 6.11).

of a cutting table 1 with a flat work surface and two movable, mutually perpendicular straightedges 2 and 3 with a sliding head 4 at the point of their intersection. The head has a scriber, or cutter, 5 that touches the table surface. A glass or plastic base plate 6 of desired size covered with a thin, dark, strippable film 7 of nitrocellulose enamel is placed on the table. By moving straightedges, each parallel to itself, it is possible to cut horizontal and vertical lines in the film, which are the outlines of the desired pattern elements. The film cuts are then peeled off the base plate to define the pattern areas.

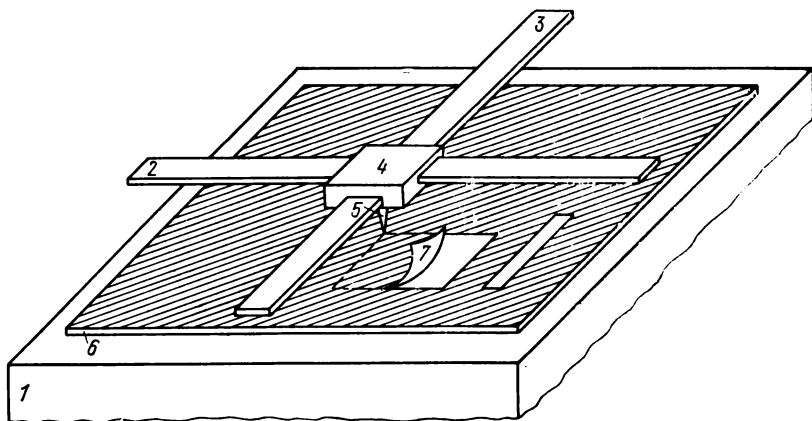


Fig. 6.14. Schematic of a coordinatograph

The next stage is a first or *intermediate reduction* of the circuit artwork, using reduction photocopiers. In the first step of photomask making, the reduction is made on a glass photoplate at the $10\times$ or $20\times$ reduction ratio. If the artwork is too large, the intermediate reproduction is made in two steps at a total reduction of $50:1$ or $100:1$.

What follows next is the *final step of photoreduction* with the simultaneous *multiplication* to produce a regular array of images on the photographic plate which then serves as a master (see Fig. 6.11). The reduction ratio used in the final step depends on the previous reductions and commonly ranges from $5:1$ to $10:1$. An array of images is produced with a *step-and-repeat* camera having an attachment for stepping the photoplate in the focal plane a small amount after each exposure. If, for example, the desired chip must measure 1.5 by 1.5 mm then the step along the horizontal and the vertical should be 2 mm.

The emulsion layer of a photomask wears away even after 15 to 20 photomasking operations. For this reason the master mask is

stored in a special room and only used whenever necessary to make copies by contact printing. The life of photomasks can be increased a hundredfold or more by metallization—replacement of the photoemulsion film by the film of a wear-resisting metal, commonly chromium. *Metallized photomasks* are produced by a photolithographic method similar to that used for metallization (see below).

Photomasks are made **in sets**, each comprising as many photomasks as there are photolithographic operations in the technological cycle. The photomasks of each set are compatible, that is, they ensure pattern alignment after bringing the registration marks into coincidence.

6.7.3. New approaches and trends. The discussed methods had long held a firm place in microelectronic technology. Even today they have not yet lost their significance. However, the trend toward an increased scale of integration and decreased size of integrated components has posed a number of problems, some of which have been solved and the others are still in the stage of studying.

Whatever the present significance of photolithography, it is not free from limitations which make themselves felt more acutely as microelectronics advances further.

One of the principal limitations relates to *resolution*, which defines the fineness of detail in the produced pattern of the mask. The fact is that the wavelengths of ultraviolet light range between 0.2 and 0.3 μm . However small the window in the photomask pattern, the size of its image in the photoresist cannot be as small as the values given above due to diffraction. The resolution of a photolithographic process is defined as 1 000 lines per millimeter (this implies the production in the photoresist of **separated** windows, or lines, 0.5 μm wide). After development and etching of the oxide, the resolution decreases to 250-500 lines per millimeter. Meanwhile, line widths and spacings of 1 or 2 μm now prove insufficiently small in the fabrication of large and superlarge ICs.

A most obvious approach to increasing resolution in lithography is to use **short-wave** radiations in exposure, for example, soft X-radiation with a wavelength of 1 or 2 nm. This approach is still in the stage of research and development since it requires the solution of a variety of complex technological problems and necessitates re-equipment to meet the demands.

Indeed, the use of short-wave radiation in itself cannot solve either the problem of decreasing the size of circuit elements or the problem of pattern alignment. What is needed here is a new technique for the fabrication of masks with submicronic pattern elements (the use of X-ray radiation is one of the approaches to the solution of the problem). Also, there is a need for new resists of increased resolution and appropriate chemicals for their treatment. Last, a basic problem that awaits its solution is to choose or develop an adequate

source of X-radiation. One of the variants is a synchrotron, an installation applied in nuclear engineering that might be adapted to fill the needs of microelectronics. However, this unique installation is too costly for use in industry on a large scale.

Even after bringing the above problems under control, it is out of the question to expect to have integrated elements with dimensions lying in the nanometer range. There are a number of factors which stand in the way of reaching this range, such as undercutting of a resist and silicon dioxide, lateral diffusion, or spreading of ions under the mask in ion implantation. Thus X-radiation lithography is expected to produce circuit elements of fractions of a micrometer in size, that is, an order of magnitude smaller than the present integrated elements. The trend aimed at decreasing the size of circuit elements, using, in particular, X-radiation and electron beams (see below), has received the name nanoelectronics (submicronic technique).

One of the weak spots in classical photolithography is mechanical contact between the photomask and substrate coated with the photoresist. This contact cannot be too perfect, so it leads to various kinds of distortion of the pattern. The competing technique is *projection photolithography* in which the pattern on the photomask is projected on the substrate with the aid of a special optical system. The technique uses an *intermediate* mask because the optical system secures the desired reduction. The regular array of images of the pattern are transferred to the substrate by shifting it stepwise under the projector.

In widespread use now are *pattern generators*. These are computer-aided automatic setups which make a photomask from nontransparent plates of various shapes according to a special program. The program provides for precision slits between these plates, which serve for exposure of the photoresist. Pattern generators do away with the complex and multistep process of making conventional photomasks.

Last years have seen the emergence of *electron beam lithography*. The essence of the technique is the following. A focused electron beam of computer-controlled intensity *scans*, line by line, the substrate surface coated with a resist. At the points which must be "exposed" the current of the beam is the highest, and at the points which must be "unexposed" the current is the smallest or equal to zero. The electron beam diameter is directly dependent on the beam current: the smaller the beam diameter, the lower the current. However, the exposure time grows with a decreasing current. Therefore an increase in resolution (decrease in the beam diameter) tends to lengthen the process. For example, the procedure of substrate scanning with a beam 0.2 to 0.5 μm in diameter can last from tens of minutes to a few hours.

One of the variants of electron beam lithography dispenses with resist masks at all and exposes the oxide layer directly to the electron

beam. At the spots of exposure the etchant removes the layer a few times faster than in the unexposed areas.

As regards the problem of pattern alignment, the trend today is to solve the problem by means of *self-alignment*. The principle underlying this trend comes to using the previously formed structural elements as masks for producing successive elements. Isoplanar technology (see Fig. 7.10) and self-aligned gate MOS transistor technology (see Figs 7.30 and 7.31) can serve as examples.

6.8. Thin Film Deposition

Thin films which lie at the root of thin-film hybrid technology also find wide use in the fabrication of semiconductor integrated circuits. Thin-film deposition techniques thus relate to general aspects of microelectronic technology.

Three basic methods are available to microelectronics for deposition of thin films on the substrate and one onto the other: *thermal (vacuum) evaporation*, *ion-plasma sputtering*, and *electrolytic (electrochemical) deposition*. There are two variants of ion-plasma sputtering: *cathode sputtering* and *ion-plasma sputtering proper*.

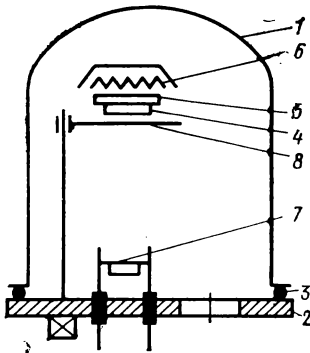


Fig. 6.15. Vacuum evaporator

6.8.1. Vacuum evaporation. Fig. 6.15 illustrates a vacuum evaporator in schematic form. This is a vacuum chamber, essentially made as a metal or glass jar 1 disposed on a bearing plate 2. An interlayer 3 placed between the jar and the plate secures the desired vacuum in the chamber after air evacuation. A substrate 4 onto which a film is to be deposited is fixed

to a holder 5. A heating element 6 is placed next to the holder to heat up the substrate during film deposition. A heater 7 serves to heat the source of a substance to be vaporized. A rotatable slide 8 cuts off the stream of the evaporant from the heater to the substrate and thus discontinues the process whenever necessary.

The heater is a filament or spiral from such a high-melting metal as tungsten or molybdenum, through which a rather heavy current is passed. The source of a substance to be evaporated can be in the form of clips hung on the filament, as a rod encircled by the spiral, as a powder poured into a crucible and caused to vaporize under spiral heat, and so on. In the last years, heating by an electron beam or laser beam has gained recognition.

The operating conditions in the chamber favor the condensation of vapors on the substrate, though a certain amount of the evaporant condenses on the jar walls. Too low a temperature of the substrate prevents a uniform distribution of the atoms being adsorbed: they form "isles" of various thickness, often without being linked together. On the contrary, too high a temperature of the substrate causes the atoms that have just arrived to "re-evaporate" and leave the substrate. Therefore, in order that the film be of high quality, the substrate temperature must lie within certain optimum limits, generally 200 to 400°C. The rate of film growth depends on many factors such as the heater temperature, substrate temperature, distance from the heater to the substrate, and the type of material being evaporated. The rate of deposition spans the range from tenths of a nanometer to tens of nanometers per second.

The strength of bonding of a film to the substrate or to another film is called *adhesion*. Some popular materials, for example gold, show poor adhesion to typical substrate materials, including silicon. Where this is the case, it is good practice to evaporate onto the substrate an *undercoat* of adequate adhesive strength and then to deposit a basic layer which adheres well to the undercoat. For example, nickel or titanium provides a good undercoat for gold.

To minimize the scattering of the evaporant atoms due to collisions with residual gas atoms, the chamber should be evacuated to a sufficiently high vacuum. A criterion for the required vacuum can be a condition at which the mean free path of atoms is a few times as large as the distance between the heater and substrate. But this condition alone does not often prove enough because any quantity of the residual gas is fraught with contamination of the film being grown and deterioration of its properties. At present, a vacuum below 10^{-6} mm Hg is considered inadequate. Some modern high-class evaporation systems operate at a vacuum of 10^{-11} mm Hg.

The discussed method is simple and gives exceptionally pure films in a high vacuum, which is its major advantage. Unfortunately it presents serious drawbacks too. The method does not allow for easy deposition of **high-melting** materials and makes it difficult if not sometimes impossible to reproduce the chemical composition of the substance being evaporated on the substrate. The latter difficulty comes from the fact that chemical compounds dissociate at high temperature and their constituents condense **separately** on the substrate. Naturally, a probability exists that a new combination of atoms on the substrate will not match the structure of the initial molecule.

6.8.2. Cathode sputtering. A scheme of this technique is presented in Fig. 6.16. Most of the construction components here are the same as in Fig. 6.15, except that the heater is absent; a cathode 6 now occu-

pies the position of the heater and performs its function. The cathode can be just the material being sputtered or a separate element brought in electric contact with the material. The substrate with its holder acts as an anode.

The chamber is first evacuated to 10^{-5} - 10^{-6} mm Hg and then a valve 8 is made open to allow a certain amount of pure neutral gas, most commonly argon, to enter the chamber, thus bringing up a pressure to 10^{-1} - 10^{-2} mm Hg. On applying a high voltage of 2 or 3 kV to the cathode (the anode being grounded for safety reasons), an **anomalous** glow discharge appears in the anode-cathode space, which entails the formation of a quasineutral electron-ion plasma.

What distinguishes an anomalous glow discharge is a strong electric field that builds up in the space near the cathode. The field accelerates positive ions of the gas so that they bombard the cathode and knock out not only electrons needed to sustain the discharge but also neutral atoms. The cathode thus gradually disintegrates. In common gas-discharge devices which build up a **normal** glow discharge, cathode destruction is impermissible, but here the knockout of atoms from the cathode is a useful process analogous to evaporation.

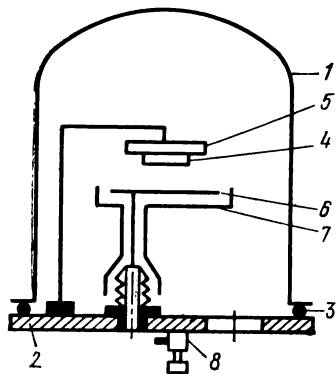


Fig. 6.16. Cathode sputtering chamber

Cathode sputtering offers an important advantage over vacuum evaporation in that the sputtering of the material used *does not require high temperature*. This process thus resolves the difficulty of depositing high-melting materials and chemical compounds.

In this method, however, the cathode (the material to be sputtered) must exhibit high conductance since it serves as an element of the gas-discharge circuit. This requirement places a limit on the list of materials to be sputtered. In particular, it proves impossible to deposit dielectrics, including many oxides and other chemical compounds being in widespread use in the technology of semiconductor devices.

Reactive, or chemical, cathode sputtering remedies the situation to a considerable degree. The method relies on the use of **active gases**, added in small amounts to the basic inert gas, which are able to form the desired chemical compounds with the cathode material being sputtered. An addition of oxygen to argon, for example, permits growing an oxide film on the substrate. Mixing argon with nitrogen or carbon monoxide can give nitrides or carbides of respective me-

tals. Depending on the partial pressure of an active gas, the chemical reaction can take place either on the cathode to yield the **ready** compound deposited on the substrate, or on the substrate—anode.

Cathode sputtering on the whole suffers from such disadvantages as relatively complex control, somewhat contaminated films and a lower deposition rate as compared with the vacuum evaporation process on account of a comparatively low vacuum built up in the chamber.

6.8.3. Ion-plasma sputtering. A scheme of the sputtering technique appears in Fig. 6.17. What distinguishes this method from cathode sputtering is an **independent**, “guarding”, discharge in the gap between an electrode 9—a *target* with the deposited material to be sputtered—and a substrate 4. The discharge known as an arc discharge takes place between electrodes 6 and 7. It features a special electron source (filament cathode 6), low operating voltages (a few tens of volts), and a high density of the electron-ion plasma. As in the cathode sputtering arrangement, the chamber here is filled with a neutral gas, but kept at a lower pressure, 10^{-3} – 10^{-4} mm Hg. B

The process of sputtering comes to the following. A negative potential of 2 or 3 kV is applied to the target with respect to the plasma (practically relative to the **grounded** anode 7). This potential is enough to produce an anomalous glow discharge causing an intensive bombardment of the target with positive ions of the plasma.

As obvious, the process does not in principle differ from cathode sputtering. The differences lie mainly in the construction of chambers. The cathode sputtering arrangement is a two-electrode setup, and the ion-plasma arrangement is a three-electrode chamber.

The process of sputtering begins with the application of voltage on the target and ends after deenergization of the chamber. A mechanical slide (see Fig. 6.15), if incorporated in the system of Fig. 6.17, enables us to achieve an additional, important end: if, prior to starting with the sputtering process, we close the slide and apply a potential on the target, **ionic cleaning** of the target will take place (see Sec. 6.6). This cleaning helps improve the quality of the film being deposited. The substrate can be cleaned in a similar manner before film deposition by applying to it a negative potential.

A positive charge stored on the target involves difficulty in sput-

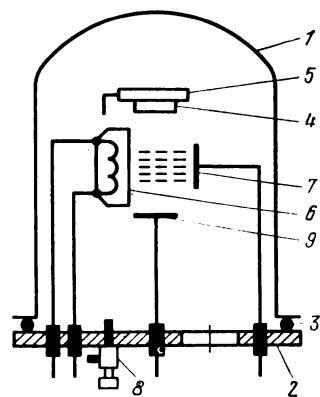


Fig. 6.17. Ion-plasma sputtering chamber

tering dielectric films since this charge impedes further ionic bombardment. This difficulty is possible to resolve in what is called *high-frequency ion-plasma sputtering*. The approach comes to applying an *ac* voltage at about 15 MHz on the target along with the *dc negative* voltage, the peak value of the *ac* voltage being slightly in excess of the *dc* voltage. The resultant voltage thus proves negative for a greater length of its period; the process of target sputtering runs as usual, and the target stores up a positive potential. For a smaller length of the period, however, the voltage is positive; then **electrons** coming from the plasma get to the target and cancel out the stored positive charge¹.

Reactive (chemical) ion-plasma sputtering offers the same possibilities of depositing oxides, nitrides, and other chemical compounds as reactive cathode sputtering described in the preceding subsection.

In comparison with cathode sputtering, ion-plasma sputtering enables a higher deposition rate and shows more flexibility (ensures ionic cleaning, permits opening the working circuit without interrupting the discharge, and so on). Besides, a higher vacuum provided in the latter process gives films of a higher quality.

6.8.4. Anodizing. This is one of the variants of reactive ion-plasma sputtering. The process involves oxidation of the surface of a metal film maintained at a positive potential with negative oxygen ions coming from the plasma of the gaseous discharge. As with purely reactive sputtering, the process requires the additions of oxygen to an inert gas. It is thus ions rather than neutral atoms that effect anodic oxidation.

In general, reactive sputtering and anodizing proceed jointly since neutral atoms and oxygen ions coexist in the plasma if it contains oxygen. In order for the anodizing process to prevail over the purely chemical sputtering, the metal film on the substrate should face in the direction opposite to the cathode to exclude neutral atoms from falling onto the film.

As the oxide layer grows, the current in the target-substrate circuit falls off since the oxide is a dielectric. To keep the current constant, it is necessary to raise the supply voltage. Because a certain amount of this voltage drops across the growing oxide film, the anodizing process takes place under the conditions at which a high strength of the field arises in the oxide film. The result is that later on, when in service, the film retains an increased electric strength.

Other advantages of anodizing include a higher rate of oxidation, since the field in the oxide film speeds up the mutual motion of metal and oxygen atoms, and a possibility of control of the process by vary-

¹ With an *ac* voltage alone applied to the target, the charge of electrons will exceed the charge of ions accumulated during the positive half-cycle since electrons have a higher mobility; the target will then be at a negative potential.

ing the current in the anode circuit. As compared to other methods, the anodizing method offers a higher quality of oxide films.

6.8.5. Electrolytic deposition. This method differs from the methods discussed above in that the working medium of the process is a liquid. But the character of the process resembles that of ion-plasma sputtering because both the plasma and electrolyte are quasineutral mixtures of ions with unionized molecules or atoms. And above all, the deposition here occurs gradually, layer by layer, as does sputtering, thereby enabling the growth of **thin** films.

Electrolytic deposition originated much earlier than any of the methods discussed, back in the 19th century. It came to be popular tens of years ago in machine-building industry for electrodepositing (nickel plating, chrome plating, and so on) of various kinds of thin coating. In microelectronics, electrolytic deposition is not an alternative of vacuum evaporation and ion-plasma sputtering; it complements each and all go together.

Electroplating depends on the electrolysis of a solution containing the ions of desired constituents. Thus the solution of blue vitriol gives a copper deposit, and the salt solution of gold or nickel provides the deposit of the respective metal.

Metal ions in a solution have a positive charge, for which reason the substrate should make a **cathode** when depositing a metal film. If the substrate is a dielectric or a low-conductivity material, it is first necessary to deposit a thin **undercoat** on the substrate by vacuum evaporation or ion-plasma sputtering and thus produce a cathode.

To effect electrochemical anodizing, the film of metal being oxidized should serve as an **anode**, and the electrolyte must contain oxygen ions.

A great advantage of electrolytic deposition over sputtering is a much higher rate of plating, the added advantage being that the plating rate is easy to control by changing the current. The electrolytic process is mainly used for depositing comparatively thick films, 10 to 20 μm and above. The quality (structure) of these films is inferior to sputtered films, but they prove quite acceptable for use in a number of applications.

6.9. Metallization

In semiconductor IC fabrication procedures, the process of metallization serves to provide ohmic contacts to semiconductor layers and also the pattern of interconnections and termination areas (contact, or bonding, pads).

The basic material used for metallization is aluminum. This metal has proved most suitable for the purpose because it features low resistivity ($1.7 \times 10^{-6} \Omega \text{ cm}$), adheres well to silicon dioxide, welds

readily to aluminum and gold wires (jumpers) in producing external leads, has high resistance to corrosion, low cost, and other attractive features.

In carrying out aluminum metallization, a uniform film of aluminum about $1\text{ }\mu\text{m}$ thick is first deposited **on the whole** of the surface of an IC (Fig. 6.18). This film comes in contact with silicon through windows 1, 2 and 3 made in the oxide layer by the preceding photomasking operation, but the main portion of the aluminum film lies on the oxide surface. After coating the aluminum film with a photoresist, exposing it through an appropriate photomask, and subsequently developing the photoresist, we obtain a **photoresist mask**.

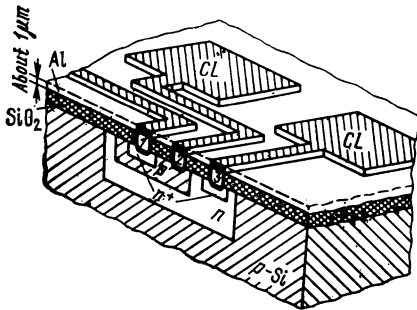


Fig. 6.18. Illustrating the process of manufacture of interconnections using the photolitho technique

Etching aluminum from the unprotected areas and then removing the photoresist gives the desired interconnection pattern with contact lands *CL*, shown in Fig. 6.18 as a hatched area adjacent to the contacts 1, 2, and 3.

The width of connection strips in conventional ICs is 10 to 15 μm , while in LSI circuits the strips are 5 μm wide and even less. The linear resistance of a strip 10 μm wide and 1 μm thick is about $2\text{ }\Omega/\text{mm}$. The contact lands for connecting metallizations to lead-out wires commonly measure $100 \times 100\text{ }\mu\text{m}$. Direct bonding of strips to external leads is impossible because the strips are very narrow.

Of course, an interconnection pattern should be free from crossings, or shortings. In ICs of a high scale of integration, however, it is impossible to design a connection layout so as to exclude crossovers, or underpasses. The common approach to solving the problem is to use a *multilayer*, or *multilevel interconnection pattern*, that is, to produce an interconnection film in a few "storeys" or levels separated by insulating layers. The required connections between various levels are made through special windows provided in insulating layers (Fig. 6.19). Isolation of conducting layers is effected by **depositing** a dielectric after completing the successive level of metallization. The dielectric used here is commonly silicon monoxide. The multilayer interconnection in modern LSI is made in two to four "storeys".

The problem involved in preparing ohmic contacts using aluminum consists in the following. An aluminum film deposited directly on the surface of silicon produces a Schottky barrier (see Sec. 3.3). This barrier in the case of an n -silicon is not ohmic but rectifying. To exclude the formation of Schottky barriers, aluminum is *fused* into silicon at about 600°C , which is close to the temperature of an Al-Si eutectic. At such a temperature, a thin layer of aluminum-silicon alloy builds up at the boundary between the aluminum film and silicon; in this layer, practically all aluminum adjacent to silicon

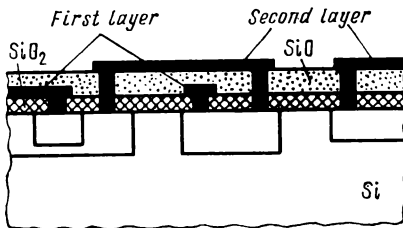


Fig. 6.19. Multilayer connection pattern

is found to be in the dissolved state. After its solidification, the compound represents an alloy of silicon with aluminum whose concentration comes close to $5 \times 10^{18} \text{ cm}^{-3}$.

Because aluminum is an **acceptor** with respect to silicon, there is a danger of the formation of pn junctions in n layers. Indeed, if the donor concentration in the n layer is below $5 \times 10^{18} \text{ cm}^{-3}$, then the aluminum atoms will produce a p -type surface layer in the n layer. To avoid this, the n layer region near the contact is doped additionally to convert it to an n^+ layer with a donor concentration of 10^{20} cm^{-3} and over (see Fig. 6.18). The aluminum concentration then proves insufficient to form the p layer, and so a pn junction does not appear.

If an n layer such as the emitter region of a transistor is heavily doped from the very beginning, there is no need for its additional doping. No problems emerge if aluminum is in contact with p layers, because aluminum dissolved in these layers tends to form p^+ -type surface regions, which improves the quality of the ohmic contact.

6.10. Assembling

After completion of the basic technological stages, metallization included, the wafer incorporating hundreds of ICs is separated into individual dice, or chips.

The operation involved in cutting the wafer into chips is called dicing, or *scribing*. The operator uses a diamond scribe to cut vertical and horizontal grooves in the gap between ICs (see Figs. 1.1 and 1.2).

He then puts the wafer on a soft rubber pad and gently breaks away the chips along the scratches in the same manner as the glazier does when he breaks up the sheet of glass along the scratch scribed by a diamond cutter. Sound chips are then mounted and encapsulated.

The assembly of integrated circuits starts with the operation called *mounting of the chip on a header*, which is the bottom of an envelope. The chip is bonded or soldered with a low-melting solder to the header in its middle portion as shown in Fig. 6.20, which illustrates a simple

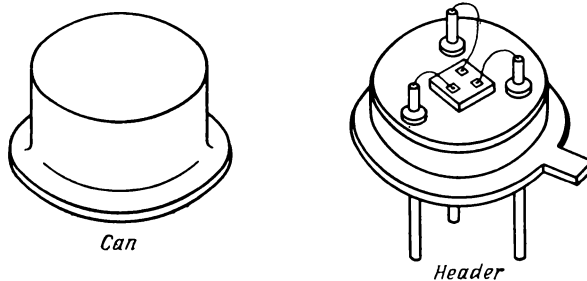


Fig. 6.20. Mounting a chip on the header

transistor. The contact lands on the chip are connected to the lead-out pins on the header with jumpers. These are fine aluminum or gold wires 20 to 30 μm thick, with one end of each wire bonded to the contact pad and the other to the end face of the pin.

Reliable contact between metal parts (here, contact of jumpers with bonding pads and lead-out pins) can be made by a variety of methods, the most popular being *thermocompression bonding*. This method uses a sufficiently high pressure to press one metal part against the other in combination with a rather high temperature, 200 to 300°C, to ensure mutual diffusion of atoms between the two parts and thus make a solid weld.

There are many variants of thermocompression bonding as regards the design of bonders, though the principle is the same. Two typical adaptations of the principle are wedge bonding and nailhead bonding (Fig. 6.21). Fig. 6.21a shows a wedge bonder, which presses a wire against the metal surface so as to produce a firm stitch bond. Fig. 6.21b shows a nailhead (ball) bonder. This is a capillary tube with a wire threaded through its internal channel. As the wire is cut off with a microtorch, a globule (ball) is formed at its end (ball formation is typical of gold wires). When the capillary is again pressed to the bonding pad, the ball spreads over, producing a contact in the form of a nailhead, hence the name nailhead bonding. The capillary is now raised to pull out a length of wire enough to connect another part. The wire is again cut off with a fine flame and the operation is repeated. Wire bonding is carried out with a microscope

if it is the operator who does the job. Automatic machines are now available to perform the operation.

After mounting the chip on the header, the final stage follows that involves *encapsulation*, or packaging, to give the device the commercial appearance. The header is connected to the can (see Fig. 6.20)

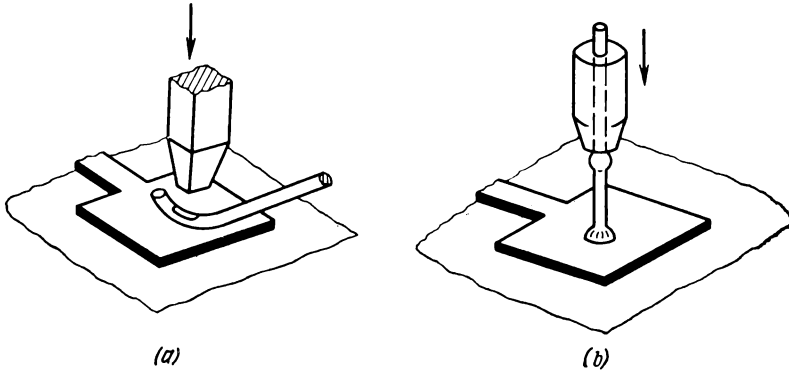


Fig. 6.21. Thermocompression bonding using blunted wedge (a) and capillary tube (b)

by hot or cold welding, the latter technique being in essence similar to thermocompression bonding. The aim of encapsulation is also to protect the chip against the influences of the environment, and there-

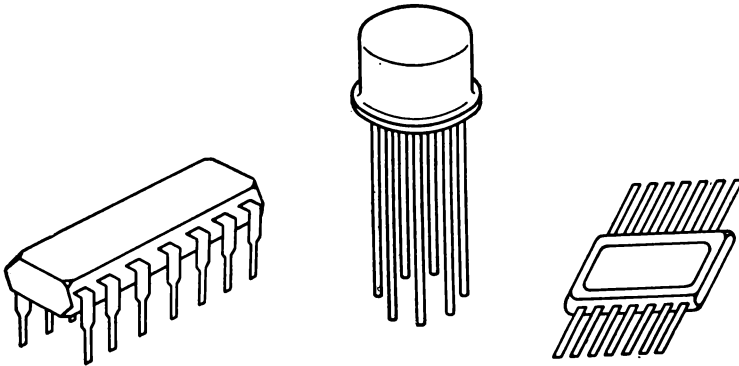


Fig. 6.22. Typical integrated-circuit packages

fore this operation necessitates the sealing in a vacuum or in a protective atmosphere such as nitrogen or argon. Subsection 6.1.2 describes the techniques of mounting uncased (unpacked) transistors on the substrate.

The main feature of assembly operations as applied to integrated circuits is the need for using multilead enclosures for ICs: small-scale ICs have 8 to 14 leads, and large-scale ICs up to 64 leads and more. Encapsulations for integrated circuits are rather diverse in shape and in other features. Along with round metal envelopes similar to the discrete transistor case shown in Fig. 6.20, in use are rectangular and flat plastic packs. Fig. 6.22 shows a plastic-encapsulated dual-in-line package, a cylindrical metal package, and a flat pack with pins running parallel or perpendicular to the enclosures. The choice of a package depends on the function of equipment and methods of its manufacture.

6.11. Thin-Film Hybrid IC Technology

According to the definition given in Sec. 1.2, the hybrid IC is a combination of passive film elements and discrete active elements. Thin-film HIC technology thus combines the thin-film passive network fabrication techniques and active component-attachment techniques.

6.11.1. Fabrication of passive elements. The techniques for fabrication of thin-film hybrid circuit elements are the same as described in Sec. 6.8, namely, vacuum evaporation and cathode or ion-sputtering of the desired material for its selective deposition onto the dielectric substrate through windows in the mask.

Thin-film hybrid technology had long used *superposed metal masks* made up of a thin bimetallic foil with openings, or windows. A layer of beryllium bronze 100 to 150 μm thick formed the base for an electrolytically deposited nickel layer 10 to 20 μm thick. The latter served to define the pattern in the mask and the former played the part of a backing.

Serious shortcomings of metal masks lie in the following. First, in deposition of films through a mask, the material being deposited also settles on the mask. This changes the thickness of the mask and gradually renders it unsuitable for further application. Second, metal masks are hardly fit for use in cathode or ion-plasma sputtering because the mask metal distorts the electric field and hence affects the rate of sputtering. In the last years the photolitho technique—the technique derived from monolithic technology—has in principle ousted the metal mask practice.

The steps of the photolithographic process here are the following. **Uniform** films of the desired materials are first deposited on the substrate to produce, for example, a resistive layer and a conductor layer on top of the resistive layer. The surface is now coated with a photoresist, and then a suitable photomask is used to define in the photoresist the desired pattern for the conductor layer, say, for con-

tact lands of a resistor (Fig. 6.23a). Next the conductor layer is etched through the windows in the photoresist and then the photoresist is removed. The contact lands on the still uniform resistive layer are now ready (Fig. 6.23b). A new layer of photoresist is deposited and another photomask is employed to produce the pattern of resistor strips (Fig. 6.23c). Etching and photoresist removal then follow to form the desired shape of the resistor with contact lands (Fig. 6.23d).

It is of course important that the etch that dissolves the conductor layer should be neutral to the resistive layer, and vice versa. There are also other limitations which we will not dwell upon here. Let us note in passing that the photomasking technique cannot give **multilayer** structures of the capacitor type. But this limitation is of little consequence since HIC technology has recently come to use essentially **discrete** capacitors as they effect economy in the substrate area.

The materials for resistive films are most commonly chromium, nichrome (80% Ni, 20% Cr), and cermet (50% Cr, 50% SiO) deposited in a vacuum. Ohmic contacts to resistive strips are accomplished in a manner shown in Fig. 6.23.

Capacitor plates are made from aluminum. Since this metal does not adhere firmly enough to the substrate, there is a need to grow first a Cr-Ti layer directly on the substrate to provide an *undercoat* for the lower plate.

As regards the requirements for permittivity ϵ , loss tangent $\tan \delta$, breakdown strength, and other parameters, the monoxides of silicon and germanium are most suitable for use as dielectric layers of film capacitors.

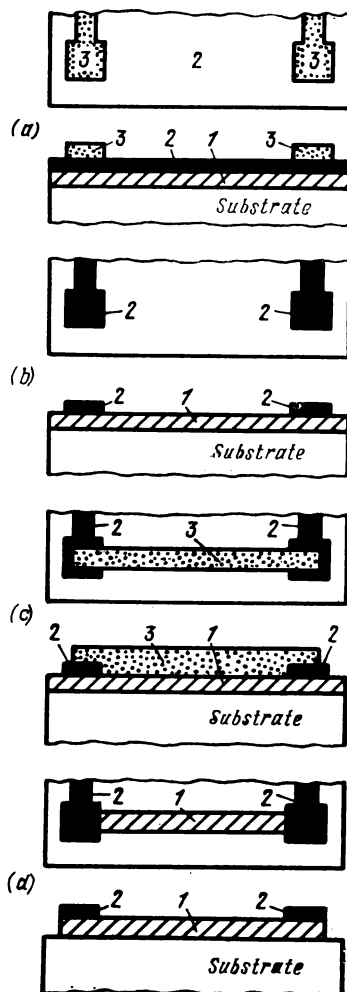


Fig. 6.23. Basic steps in the fabrication of thin-film resistors using a photomasking technique (a) photoresist mask 3 for pattern of conductor layer 2; (b) ready conductor pattern 2; (c) photoresist mask 3 for pattern of resistive layer 1; (d) ready resistor with leads

Oxides Ta_2O_5 and Al_2O_3 hold a specific position among dielectrics. They are obtained by anodic oxidation of lower titanium or aluminum plates rather than by deposition.

The materials for conductive films and ohmic contacts are as a rule gold with a Cr-Ti undercoat or copper with a vanadium undercoat. Conductive films and contact pads are usually 0.5 to 1 μm thick. Contact pads measure from 200 by 250 μm and over.

The thickness of films being grown is controlled by a number of methods. One of them, applicable only for resistive films, uses a *monitor*. This is an auxiliary layer of a given geometry located at the substrate periphery, deposited simultaneously with the main pattern, and fitted with two external leads for connection to an ohmmeter. When the resistance of the monitor reaches the value corresponding to the desired film thickness, the evaporation is terminated by shutting off the flow of evaporant with a rotatable slide.

In another control method, the monitor employed is a thin quartz plate connected via external leads to the resonant circuit of an oscillator. As known, a quartz plate displays the properties of an oscillatory circuit and has a resonant frequency that changes with the plate thickness in a unique manner. During evaporation the plate thickness grows, and so the oscillator frequency changes. Frequency variations are easy to measure and thus determine the right moment when it is necessary to terminate the deposition.

The substrates of thin film hybrids must primarily have good insulating properties. Besides, it is desirable that they show a low dielectric permittivity, high thermal conductivity, and sufficient mechanical strength. The substrates must match closely the deposited films in the TC of thermal expansion. The typical parameters of substrates are the following: $\rho = 10^{14} \Omega \text{ cm}$, $\epsilon = 5$ to 15, $\tan \delta = 2 \times 10^{-4}$ to 20×10^{-4} , and $TCL = 5 \times 10^{-6}$ to 7×10^{-6} .

At present, *glazed* and *ceramic* substrates are most popular; glass has lost its former importance. Glazed ceramics are crystal modification of glasses (common glass is amorphous) and ceramics are mixtures of oxides in vitreous and crystalline phases, the main constituents being Al_2O_3 and SiO_2 .

The thickness of substrates averages 0.5 to 1 mm depending on the area. The area of thin-film substrates far surpasses the area of chips for semiconductor ICs. The standard dimensions of substrates range from $12 \times 10 \text{ mm}$ to $48 \times 30 \text{ mm}$. The requirements for surface finish are approximately the same as they are for silicon: the permissible roughness does not exceed $25 \times 50 \text{ nm}$.

Hybrid ICs, like semiconductor ICs, are generally fabricated by batch technology on a common glazed or any other substrate of a large area. After the passive circuitry and metallization are complete, the substrate is scribed in the same manner as a semiconductor IC slice to separate it into individual circuits. Active components are

then mounted and bonded to the circuit, and the complete hybrid is enclosed in a case.

6.11.2. Component attachment. The discrete devices to be mounted on the passive circuitry are unpackaged diodes and transistors. A simple variant of the unpackaged transistor is a chip separated from the slice by scribing; the chip is provided with fine wire leads bonded to its three contact pads and protected from the environment by a drop of epoxy that envelopes the chip on all the sides. The transistor is attached to the substrate near the elements to which it has to be connected (see Fig. 1.5), and then its lead wires are bonded by thermocompression to the respective contact lands on the substrate.

There are two more types of unpackaged transistor whose assembly technique differs from the described above. The first type is a *ball-lead* transistor (Fig. 6.24a). Balls 50 to 100 μm in diameter from gold, copper, or Sn-Sb alloy are formed on the bonding pads of the transistor and columns of the same material, 10 to 15 μm in height and 150 to 200 μm in diameter, are then built up on the bonding pads of the substrate, (Fig. 6.24b). The balls and columns must register exactly when brought into contact. The method of bonding the chip to the substrate is known as a *flip-chip method*: the chip is mounted, face down, on to the substrate so that the balls make contact with the solder columns (Fig. 6.24c). A pressure applied to the chip in combination with elevated temperature, which is in essence thermocompression, provides for reliable soldered joints. The flip-chip, or face-down, bonding is a *multiple-joint* (multiple attachment) technique: the single face-down operation

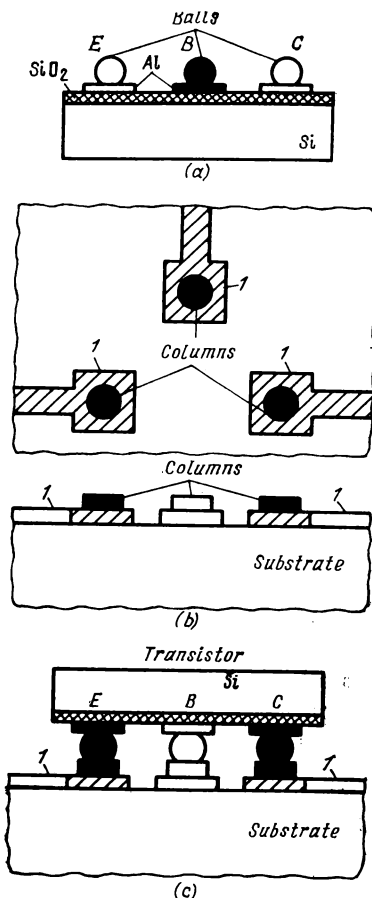


Fig. 6.24. Steps in mounting an uncased ball-lead transistor on the substrate

(a) transistor with ball leads; (b) column on the substrate of film IC; (c) connection of balls to contact columns; 1—bonding pads with leads on the substrate

(multiple attachment) technique: the single face-down operation

gives the three required connections. In comparison with wire bonding, the flip-chip approach reduces the number of connections by one half. Besides, the transistor, as a flip-chip, does not need a special area on the substrate. The main problem involved in this assembly technique is the difficulty of ensuring proper registration of balls and columns since the chip is face down and does not permit the operator to view the areas of contact.

The *beam-lead* technology obviates the difficulty of alignment of contact areas. The method is applicable to the second type of unpackaged transistor, known as a beam-lead transistor (Fig. 6.25a).

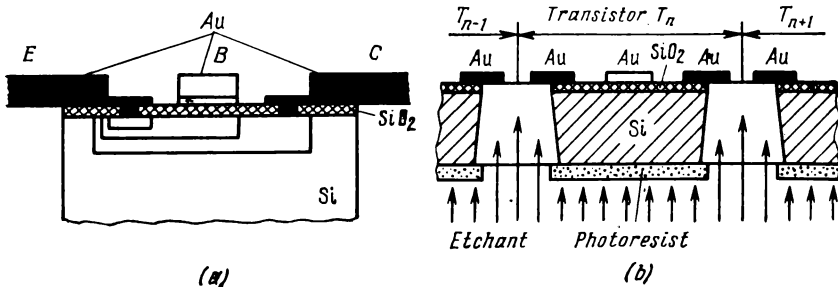


Fig. 6.25. Unpackaged beam-lead transistors

(a) beam-lead transistor; (b) fabrication of beams and separation of transistors arranged on the slice

Here the contact pads extend beyond the finished chip: they project 100 to 150 μm beyond the chip edge and hang over it as beams; hence the name beam leads. The thickness of beams is 10 to 15 μm , much greater than the thickness of metallizations on the chip. That is why beams are produced not by evaporation but by electrolytic deposition of gold with a titanium underlayer. The length of beams with projections is 200 to 250 μm , the width being the same as that of common contact lands, 50 to 200 μm .

The beam-lead pattern fabrication procedure involves **through-etching** of silicon through the windows in the photoresist mask deposited onto the back of the slice (Fig. 6.25b). Along with the fabrication of a set of beam leads, through-etching enables separation of the slice into individual chips without resorting to the conventional scribe-and-break technique. Before starting the etching, the slice is glued, with its face upward, to a sheet of glass and then thinned by lapping off its back to reduce its thickness from 200–300 μm to 50 μm . This operation cuts down the etching time and excludes lateral etching of the slice. After the slice is etched through along with the wax, the chips are readily separated from the glass.

The active components can be facedown-bonded to beam leads in the same way as flip-chip components. Alignment of beam leads to

contact pads on the substrate does not present difficulty since the beams extend far over the edges of the chip. Faceup bonding is also possible, but then the substrate must have a groove for the chip.

Though the fabrication of ball and beam terminals requires more skill and money than that of wire leads, the former allows for much simpler and cheaper assembly operations, which generally account for the largest share of the finished product cost, and enable a noticeably increased yield of ICs and improved reliability.

6.12. Thick-Film Hybrid IC Technology

Passive elements of thick-film HICs are produced by **selective** deposition of semiliquid pastes—*vitreous enamels*—onto the substrate with the subsequent drying and firing to allow the paste to fuse into the substrate. The films acquire the desired thickness **at once**, but not gradually, layer by layer, as is the case with thin-film technology.

The basic steps involved in the fabrication of thick films are as follows:

- (a) deposition of a layer of paste onto the substrate through a detachable mask, known as a stencil screen, hence the name *screen printing technique*;
- (b) drying at 300 to 400°C for evaporation of the solvent and conversion of the paste from the semiliquid to the solid state;
- (c) firing of the solidified paste on to the substrate at 500 to 700°C depending on the paste composition.

Firing is the most critical operation in the technological cycle and requires an accurate control of temperature to within $\pm 1^\circ\text{C}$.

A major constituent of all the pastes is the glass *frit*—a finest glass powder—which serves as the base for resistive, conductive, or dielectric powders to form the pastes of the desired composition, depending on the function they must perform. A **disperse**, quite homogeneous mixture of the frit and an additive acquires a viscosity after introducing special organic substances and solvents into the mixture. At the stage of drying the solvent vaporizes, and the organic substances bond the powder particles, producing a homogeneous compact mass.

Silver or gold commonly serves as an admixture for conductive pastes, silver and palladium taken in the proportion 1 to 1 for resistive pastes, and barium titanate of high permittivity for dielectric pastes. Selecting the constituents of the composition and the percentage of each permits changing the electrical characteristics of films over a wide range (see Sec. 7.11).

The masks for deposition of pastes onto the substrate are mesh screens (Fig. 6.26a). These are the fine woven mesh from capron or

stainless steel wire attached to the bottom of a screen frame¹. The screen mesh size is about 100 μm , and the diameter of filaments is about 50 μm . The mesh has its larger area coated with a film called a stencil; the windows define the desired pattern which is produced by etching the openings in the film using the photomasking technique. Because of the mesh nature of the screen, the pattern openings smaller than 100-200 μm are difficult to obtain. This places the limit on a minimum size of thick-film hybrid elements and on the line width.

For transferring the desired pattern to a substrate, the frame carrying a stencil screen is filled with a paste and placed 0.5 to 1 mm

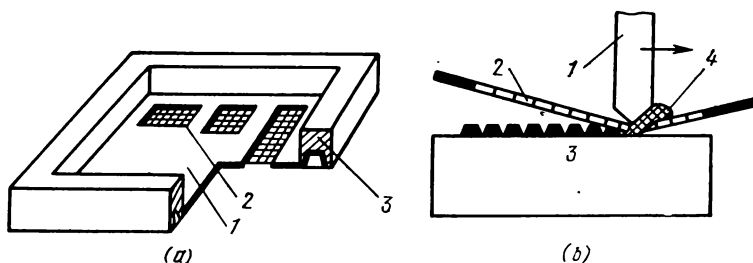


Fig. 6.26. Selective deposition of paste into the substrate

(a) stencil screen; 1—stencil; 2—window; 3—screen frame; (b) squeezing paste through the screen; 1—squeegee; 2—screen; 3—substrate; 4—paste

above the substrate. A special knife, called a *squeegee*, is then lowered onto the screen to press the paste down into openings as its blade travels along the frame (Fig. 6.26b). Despite the fact that the squeezing approach seems simple, the operation itself calls for precision; the quality of the film being deposited and the repeatability of the parameters depend on the bevel angle of the blade and its inclination to the substrate, the squeegee speed and pressure and on other factors.

Generally speaking, the mesh for a stencil screen is not an obligatory part: it is quite possible to squeeze the paste through mesh-free windows. But the quality of films obtained without the mesh is poorer. The reason is that the mesh ensures more homogeneous layers because of better merging of individual “droplets” while they pass through mesh openings. The thickness of deposited films depends on the diameter of threads and mesh size, and commonly ranges from 20 to 40 μm .

The substrates for thick-film HICs must generally meet the same requirements as those placed on thin-film hybrid circuit substrates.

¹ The former material used as a screen fabric was silk, for which reason the method of depositing pastes through woven silk screens was often referred to as *silk-screen printing* (silk screening).

The thermal conductivity of substrates often requires particular consideration since the thick-film variant of HICs is typical of relatively high-power circuits. The widespread materials for thick-film substrates are high-alumina ceramics (96% Al_2O_3) and high-berillia ceramics (99.5% BeO); the latter have 7 to 10 times the thermal conductivity of the former, but are inferior to alumina ceramics in mechanical properties. An important distinguishing property of substrates for thick-film hybrid circuits is that they must have a **sufficiently rough** surface to provide the required adherence of the paste to the substrate. The degree of roughness is defined by surface unevennesses of up to 1 or 2 μm .

The methods of attachment of active components to thick-film networks are the same as for thin-film hybrid circuits, but the dimensions of contact lands are made larger, 400 by 400 μm .

On the whole, the thick-film hybrid approach is attractive for its simplicity and has the added advantage of low product cost. But in comparison to thin-film technology, the package density of thick-film hybrids is lower because of the larger width of lines, and the spread in parameters is greater because the thickness of films is difficult to control.

7.1. General

Let us recall that *integrated elements* are **inseparable** constituent parts of an integrated circuit, be it a semiconductor or hybrid IC; they cannot be specified separately and supplied as individual circuit components. One of the features of integrated circuit elements, which distinguishes them from analogous discrete devices consists in that they have *conductive and capacitive coupling with the common substrate* and, sometimes, with each other. For this reason mathematical and physical models (equivalent circuits) of integrated elements differ somewhat from the models of discrete analogs.

A second important feature of integrated elements is that in comparison with discrete devices the elements of ICs are made *in a single manufacturing process*. For example, all resistors of a film IC are produced at one time, and hence have the same thickness and resistivity. They can only differ in length and layer width. As regards a semiconductor integrated circuit, the working layer of a resistor is deposited at the same time as the base layer of a transistor, and so both layers have the same electrophysical parameters. In other words, *there are a fewer "degrees of freedom" in the manufacture of integrated elements than is the case in the manufacture of discrete analogs*. As a rule, it is possible to vary only the surface **geometry** of integrated elements, that is, to alter the length and width of elements rather than the depth of layers and their electrophysical parameters. As a result, the *parameters of integrated elements are essentially correlated* (interrelated) and their values are *limited*, which is not the case with discrete components.

Along with the above features, it is worth noting that the progress of microelectronics has led to the appearance of integrated elements which have no analogs in discrete electronics. These are multiemitter and multicollector transistors, Schottky-barrier transistors, and others. The traditional components such as diodes and capacitors have changed in design, and the range of their parameters have changed too. In semiconductor ICs there are no analogs of such traditional components as inductors, to say nothing of transformers.

Integrated components, as noted in Ch. 1, are such constituent parts of hybrid microcircuits that can be specified separately and supplied as individual products. The components of a hybrid IC are add-on devices which differ from "common" discretized only in constructional form. Diode and transistor chips can serve as an example.

The main elements of bipolar semiconductor ICs are *npn* transistors. It is these devices that make a guideline in the development of new technological cycles: a major aim the technologist has to strive for here is to ensure the optimal parameters of these devices. The technology of all other elements such as *pnp* transistors, diodes, and resistors must “adapt itself”, as it were, to the technology of *npn* transistors. This “adaptation” means that the process of manufacture of other elements should avoid, where possible, **additional** operations; it is desirable that the same working layers such as collector, base, and emitter regions can be tailored to perform other functions. This explains the coinage of such phrases as “the base layer serves as a resistor” or “the working layer of a resistor results from base diffusion”.

Until recently, the main elements of MOS circuits were induced *p*-channel MOS transistors. They determined the scope and outlines of the technological cycle, which in turn served as a guiding line for the technology of other elements. Last years have seen the appearance of high-quality *n*-channel MOS transistors after industry has managed to overcome certain manufacturing difficulties. These transistors tend to occupy a leading place in MOS transistor technology.

7.2. Isolation of Circuit Elements

Figure 7.1 shows two *npn* transistors and a diode formed in the common *n*-type silicon substrate. The collectors of both transistors and the cathode of the diode are seen to be electrically coupled together

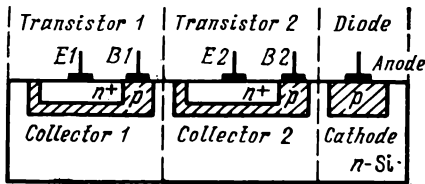


Fig. 7.1. Conductive coupling of bipolar integrated elements through the substrate in the absence of isolation

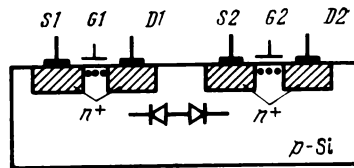


Fig. 7.2. Isolated elements of a MOS IC

through the substrate. Such couplings are as a rule objectionable: they do not correspond to the desired circuit configuration. Hence, *the elements of bipolar semiconductor ICs need to be isolated from one another in order that the necessary connections might be accomplished only by metallization.*

In MOS transistor ICs, the sources and drains of adjacent transistors are separated by reverse-biased *pn* junctions (Fig. 7.2). The conductive coupling between the adjacent elements results only from

a negligibly small leakage current of the reverse-biased junctions. This coupling can generally be neglected. As for the exchange of carriers which form conducting channels, this can only take place at distances smaller than 5 to 10 μm . Such small distances between elements are not specific to modern integrated circuits, excepting charged-coupled devices (CCD) described in Sec. 10.9.

So the elements of MOS circuits do not generally require isolation. MOS transistors can thus be located close to each other to increase the packing density and save space on the chip. This is one of the important advantages of MOS transistor integrated circuits over bipolar transistor ICs.

7.2.1. Comparative estimation of isolation methods. All the known isolation methods can be divided into two main types: *pn junction*

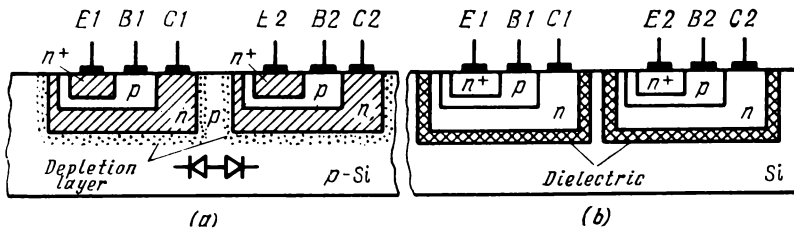


Fig. 7.3. Basic isolation techniques for integrated elements
(a) *pn* junction isolation; (b) oxide isolation

isolation (diode isolation) and oxide (or dielectric) isolation. Both types of isolation are illustrated in Fig. 7.3.

A depletion layer of the *pn* junction has a very high resistivity close to that of some dielectrics, especially at a heavy reverse bias. That is why the two types of insulation specified above differ not so much in the resistivity of the insulating layer as in the **structure** of this layer. The *pn* junction insulation approach belongs to **single-phase** methods because the material (silicon) on both sides of and within the insulating layer is the same. The oxide insulation approach relates to **two-phase** methods because the material (phase) of the insulating region differs from the substrate material (silicon).

From Fig. 7.3a it is obvious that *the method of pn junction insulation comes to building up two reverse-biased diodes between the adjacent elements.* The approach is similar to that employed for isolation of MOS circuit elements (see Fig. 7.2). So that both insulating diodes will be under reverse bias (regardless of potentials on the collectors), the substrate should be *at a maximum negative potential* applied from the IC power source. The same approach also holds for *n*-MOS transistor ICs.

The pn junction isolation technique is well compatible to the general process of bipolar IC fabrication. The limitations of this technique are reverse currents in pn junctions and barrier capacitances.

Dielectric isolation is essentially more perfect and efficient (Fig. 7.3b). At room temperature, the leakage currents in a dielectric are 3 to 5 orders of magnitude smaller than the reverse currents in pn junctions. As regards parasitic capacitance, dielectric isolation cannot obviously eliminate it completely too. But it is easy to reduce

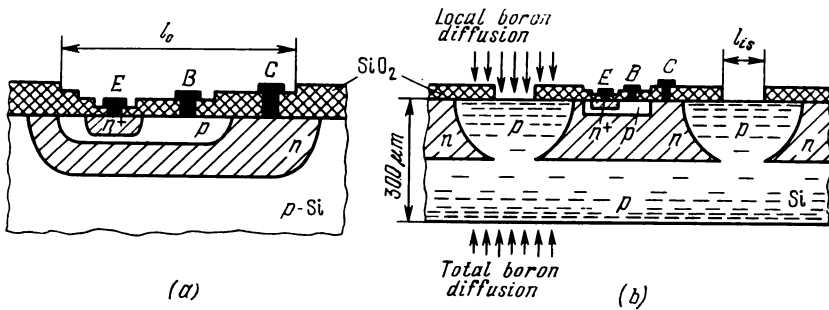


Fig. 7.4. Possible pn junction isolation techniques

(a) diffused-collector technique: l_c is the dimension of window for collector diffusion; (b) triple or two-way diffusion technique: l_{is} is the dimension of window for isolation diffusion

this capacitance below the barrier capacitance by choosing a material of low permittivity and increasing the dielectric thickness. In dielectric isolation, the parasitic capacitance is generally an order of magnitude smaller than in pn junction isolation.

One more important advantage of dielectric isolation is the possibility of **selective** doping with gold. In a single-phase system (using pn junction isolation), **local** doping is impossible: gold spreads over the entire slice because it has a high diffusion coefficient (see Fig. 6.8). In a two-phase system (using dielectric isolation), there is a possibility of diffusing gold into the islands where it is desirable to decrease the carrier lifetime, leaving intact adjacent islands which do not require doping. This possibility stems from the fact that gold diffuses into silicon much faster than into dielectrics.

Despite its advantages, dielectric isolation has not ousted pn junction isolation because it calls for a rather complex technological process involving additional operations intended to form a "foreign" dielectric layer.

7.2.2. PN junction isolation. Various methods are available for growing isolation junctions. Thus purely planar technology formerly used the methods of *diffused-collector* (Fig. 7.4a) and *triple*

diffusion, or two-way diffusion (Fig. 7.4b). Both of these methods suffer from serious disadvantages.

In the structure of Fig. 7.4a, the collector n -type layer being grown at the stage of first diffusion is **inhomogeneous**: the impurity concentration grows from the bottom toward the surface, so at the interface of the base region this concentration is rather high and hence the breakdown voltage of the collector junction is comparatively low. Besides, the diffused-collector process itself is rather complex.

In the structure of Fig. 7.4b, the isolation of elements is accomplished by total diffusion of an acceptor impurity through the **back** of the n -type slice and by **local** diffusion of the same impurity through the **windows** in the **face** of the slice. The depth of either diffusion is equal to half the slice thickness so that one diffusion region links with the other. In the upper part of the slice there appear the **islands** of the starting n -type silicon, which are the collector regions for the transistors to be formed. In distinction to the previous method,

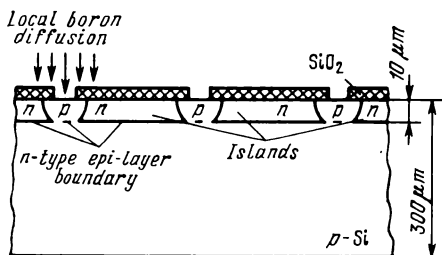


Fig. 7.5. Basic pn junction isolation techniques for planar-epitaxial ICs

the triple diffusion method provides for a **homogeneous** collector region. A major disadvantage of this method is the necessity of carrying out a very **deep** diffusion to grow a layer 100 to 150 μm in depth. The diffusion process takes 2 or 3 days and more, which makes the method uneconomical. Moreover, lateral diffusion (see Fig. 6.5b) tends to extend the insulating p type layers on the slice surface; the length of the layers becomes equal to about the slice thickness, that is, exceeds the dimensions of ordinary transistors. The space utilization factor thus decreases substantially.

A recent approach comes to dispensing with a monolithic n -type slice and using instead a thin n -type epitaxial layer grown on the p -type substrate (Fig. 7.5). The problem of isolation becomes substantially simpler: the diffusion, called *isolation diffusion*, which enables the formation of collector islands is performed *only through the upper surface* of the slice to a depth equal to the epitaxial layer thickness, commonly not over 10 to 15 μm . Thus the time of diffusion does not exceed 4 to 6 h, and the expansion of insulating p -type layers due to the sideways diffusion is merely fractions of that inherent in the two-way diffusion method (see Fig. 7.4b). The value of the chip space utilization factor is quite acceptable.

The n -type islands left in the epitaxial layer after completion of isolating diffusion are used in the subsequent stages of the manu-

facturing process for obtaining the desired integrated elements, primarily transistors¹.

The simplest islands shown in Fig. 7.5 find limited uses. The transistors formed in these islands (Fig. 7.6a) show a high series collector resistance r_{sc} , 100 Ω and above. A decrease in the resistivity of the epitaxial layer does not remedy the situation since this reduces the breakdown voltage of the collector junction and raises the collector capacitance. A more rational and typical approach is to use a so-called *buried n^+* layer located at the island bottom. The role of such a low-resistance layer is obvious from the structure shown in Fig. 7.6b.

Buried layers are diffused into the slice before growing the epitaxial layer. During the epitaxial process the donor atoms of the buried layer diffuse under high temperature into the growing n -type epitaxial layer. The buried layer thus partially shifts into the epitaxial layer so that the well bottom becomes "raised" a few micrometers in this region. An excessive diffusion of donors from the buried layer into the epi-layer can cause the buried n^+ layer to contact the base p layer. To prevent this, a diffusion chosen for the buried layer should have a rather small diffusion coefficient. This can be antimony or arsenic.

The buried n^+ layer not only decreases the series collector resistance, which is its primary function, but also offers some other advantages. Thus it raises the inverse gain of a transistor and diminishes the excess charge in the collector region that builds up in the double injection mode.

The epitaxial-diffusion method is at present the simplest and widespread variant of the pn junction isolation technique. In use are also more complex variants of this technique, one of which is *collector diffusion isolation* (CDI) illustrated in Fig. 7.7.

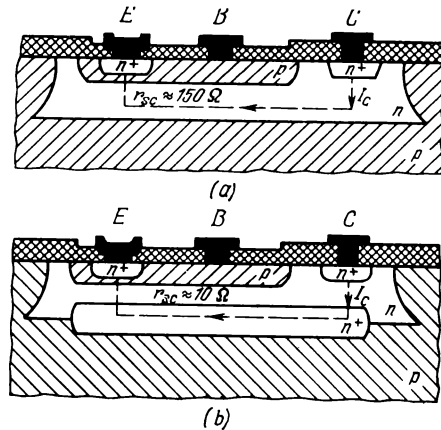


Fig. 7.6. Structure of an n pn integrated transistor
(a) without buried layer; (b) with n^+ buried layer

¹ The n^+ layer grown at the same time as the emitter n^+ layer and located under the collector electrode prevents the formation of a parasitic pn junction when firing aluminum into the n layer (see p. 204).

In the CDI, the epitaxial layer, 2 or 3 μm thick, is of the p -type conductivity. The buried n^+ layers are grown beforehand. The isolation diffusion process employs a **donor** impurity (phosphor); the diffusion depth is equal to the distance from the surface to the buried layer. The diffused regions divide the layer on the slice into p -type islands for base layers to be formed, while the n^+ layer together with isolating n^+ diffusions forms the collector region. The isolating diffusions here perform a useful function and thus do not affect the space

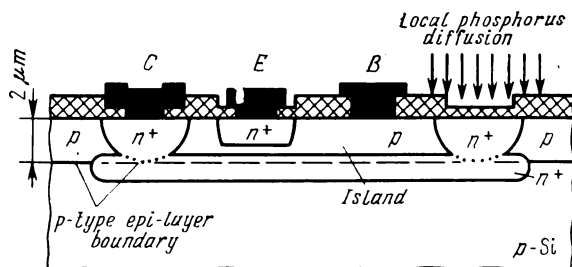


Fig. 7.7. Collector diffusion isolation technique

utilization factor. With the CDI variant employed, this factor is much higher than when using the main variant of the isolation technique (see Fig. 7.5). Because of the high impurity concentration in n^+ layers, however, the CDI method offers lower breakdown voltages for the collector junction and gives higher values of collector capacitance. Besides, the method calls for an additional diffusion of an acceptor impurity into p -type islands to render the base inhomogeneous and thus produce in it an internal accelerating field.

7.2.3. Dielectric isolation. Historically, the first method of isolation by a dielectric was an EPIC (epitaxial passivated IC) process. The steps of this process are shown in Fig. 7.8. The original slice of n -type silicon is coated with an n^+ -type epitaxial layer, 2 or 3 μm thick (Fig. 7.8a). Next, grooves 10 to 15 μm deep are etched in the slice through a mask, and then the whole of the surface is oxidized (Fig. 7.8b). Let us note that both isotropic and anisotropic etching can be used here (see Sec. 6.6), the latter being considered in Subsec. 7.2.4. The oxidation being over, a layer of polycrystalline silicon, 200 to 300 μm thick, is deposited on the surface. Further, the slice is inverted and lapped off as far as the bottom of the channels to form isolated n -type islands with a buried n^+ -type layer in polysilicon (Fig. 7.8d). The SiO_2 layer now isolates the circuit elements (compare with Fig. 7.3b). The main difficulty encountered in the EPIC process is the precision lapping of the single crystal slice. With a thickness of the layer to be removed averaging 200 to 300 μm ,

the error of lapping over the entire surface must be within 1 or 2 μm .

If at the second stage of the process (see Fig. 7.8c) a dielectric (ceramic) layer is deposited instead of a poly-Si layer, the *ceramic type of isolation* results. This variant offers a better resistive and capacitive decoupling of elements, but is more complex and expensive.

In widespread use now is the technique known as SOS (*silicon-on-sapphire*). The basic steps of the process are shown in Fig. 7.9. Sapphire has basically the same structure of its crystal lattice as silicon. This makes it possible to grow an epitaxial layer of silicon (Fig. 7.9a) on the sapphire slice and then etch through this layer as far as sapphire to form silicon islands for subsequent realization of integrated elements (Fig. 7.9b). The dielectric sapphire isolates the islands on the underside and the air does so on the sides. That is why the SOSIC isolation technique is often placed into the class of *air isolation*. A serious drawback of this technique is the irregular surface which causes difficulties in depositing the metal interconnection pattern.

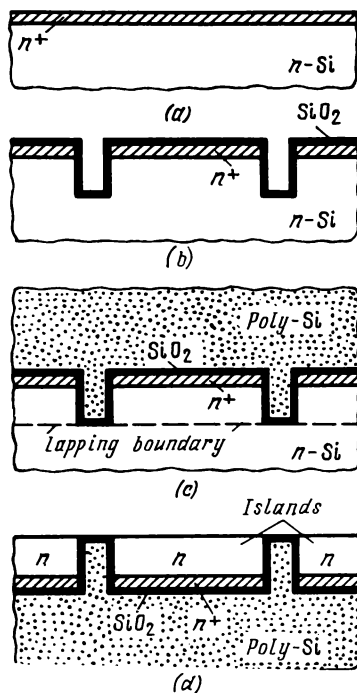


Fig. 7.8. Oxide isolation technique (EPIC process)

(a) original structure; (b) etching of grooves and oxidation; (c) poly-Si deposition; (d) finished structure showing islands with n^+ buried layer

7.2.4. Combined methods of isolation. A recent isolation technique that enjoys widespread use is the *isoplanar process*. The principle of the process lies in local **through-oxidation** of an epitaxially grown n -type silicon layer 2 or 3 μm thick (Fig. 7.10). The locally oxidized regions divide the epitaxial layer into n -type islands. The result is similar to that obtained in the isolation diffusion of Fig. 7.5. But here the isolating side walls of islands are dielectric (oxide) rather than semiconductor layers. The island bottoms, however, are isolated by the reverse-biased pn junctions. This feature explains why the isoplanar process belongs to the combined methods.

Each island is in turn divided by the oxide into two parts 1 and 2 as shown in Fig. 7.10a. In the main part 1 the base and emitter regions of a transistor are formed, and in the second part 2 an ohmic contact of the collector is produced (Fig. 7.10b). Both parts are linked

together via the buried n^+ layer. This layout does away with all the four side (vertical) walls of the collector junction, thereby reducing the collector capacitance.

Local oxidation of the epitaxial layer cannot be achieved through an oxide mask. This is because the mask will change its geometry

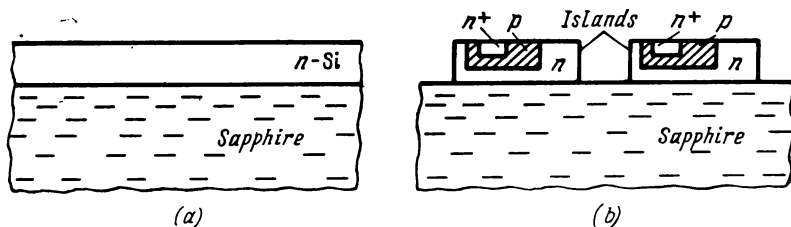


Fig. 7.9. Silicon-on-sapphire technique

(a) original structure; (b) islands grown

during through-oxidation of the epitaxial layer. For this reason the isoplanar process uses silicon nitride masks.

As compared with the classical method of isolation diffusion, the isoplanar method affords a greater packing density of elements (more

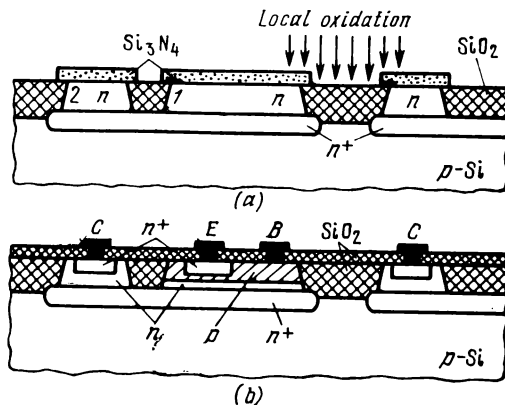


Fig. 7.10. Isoplanar technique

(a) structure prior to base diffusion; (b) finished transistor structure

efficient use of the chip area) and also ensures better frequency and transient response of a transistor.

A *V-groove isolation method* is illustrated in Fig. 7.11. Instead of through-oxidation of the epitaxial layer, this method uses anisotropic etching to etch through the layer (see Sec. 6.6). The crystal surface must have the (100) orientation to ensure etching along the

(111) planes as shown in Fig. 6.10b. The windows in the mask are so dimensioned that the (111) faces “converge” just a little below the boundary of the epitaxial layer and form V-shaped grooves (hence the name of the method). The relation between the width and depth of the groove is strictly definite: $l/d = \sqrt{2}$. At a depth of 4 or 5 μm , the groove width will be merely 6 or 7 μm , so the loss in area for the

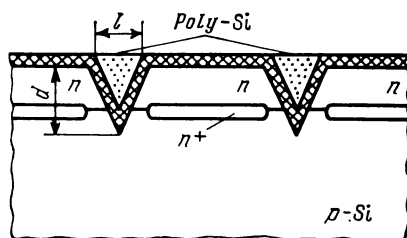


Fig. 7.11. V-groove isolation technique

isolation is rather insignificant. A limitation of the method is the necessity of using the (100) plane which shows an increased density of surface defects (see Fig. 2.5).

After its etching, the surface is oxidized as is done in the EPIC process. But in distinction to the EPIC process, here the subsequent growth of the polysilicon layer is only aimed at leveling off the surface to facilitate metallization. For this, it is enough to fill up the grooves only.

7.3. NPN Transistors

Since *n**p**n* transistors are the basic elements of bipolar ICs, we shall discuss them in greater detail along with the manufacturing techniques. We assume that the manufacturing process essentially employs isolation diffusion. Some features of other isolation techniques will be mentioned where necessary.

7.3.1. Impurity distribution. Fig. 7.12 shows the distribution of impurities in the layers of an integrated transistor with a buried *n*⁺ layer (see Fig. 7.6b). It should be pointed out that the distribution of the **effective** concentration of acceptors in the base layer is **nonmonotonic**, and hence the hole distribution is nonmonotonic too. On the right of the maximum the hole concentration gradient is negative, and the built-in field is accelerating with respect to injected electrons (see p. 75). This is an inherent feature of all drift transistors. But on the left of the maximum the concentration gradient is positive, and hence the field is retarding. *A region with the retarding field causes a certain increase in the resultant transit time of carriers that*

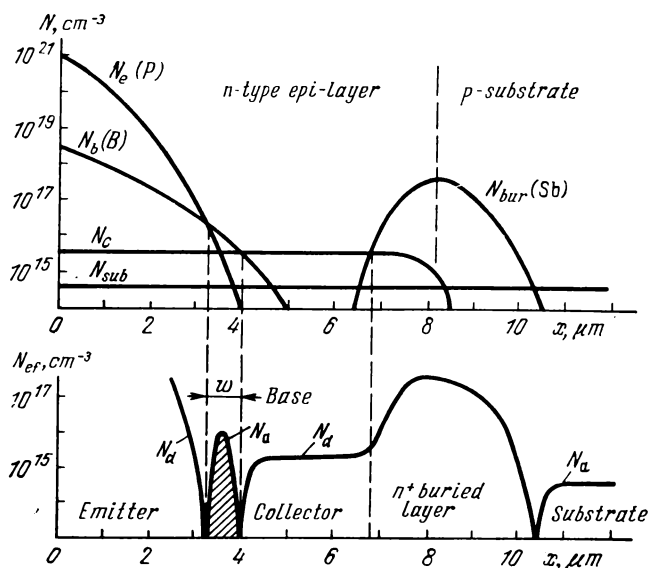


Fig. 7.12. Distribution of impurity concentration in the structure of an $n p n$ transistor and distribution of effective concentrations

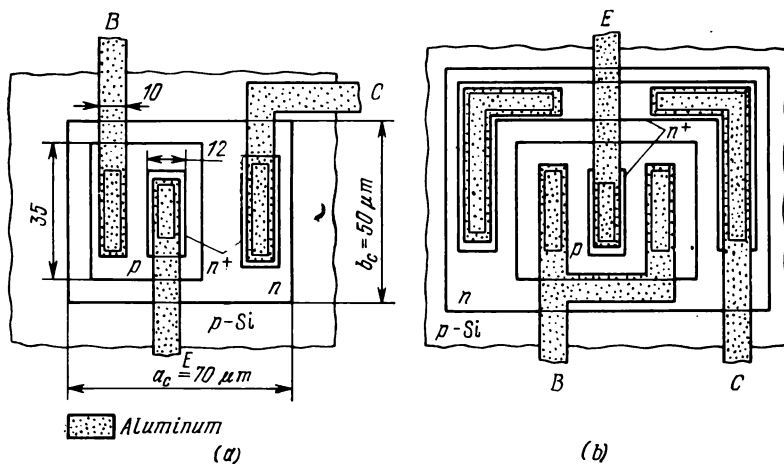


Fig. 7.13. Plan geometry of transistors
(a) asymmetric; (b) symmetric

cross the base. But calculations show that this increase is in the order of merely 20 to 30% and thus can be disregarded in approximate estimates.

7.3.2. Geometry and performance parameters. The plan geometry (topology) of integrated transistors can be of a few variants. Two of them are illustrated in Fig. 7.13.

The first variant of topology (Fig. 7.13a) is called **asymmetric**: the collector current flows to the emitter only in one direction, from the right as shown in the figure. The second variant (Fig. 7.13b) is called **symmetric**: the collector current flows to the emitter from three sides. Here the series collector resistance r_{sc} is thus only one-third that for the asymmetric layout.

Another feature of the second transistor geometry is that the contact window and the collector metallization are divided into two portions. Such a design makes it easier to fabricate the interconnection pattern: the aluminum stripe, for example, the emitter finger shown in Fig. 7.13b, can lie above the collector on the protective oxide covering the surface of an IC (for more detail, see Subsec. 7.9.1).

Figure 7.13a gives the typical dimensions of *npn* transistor layers; Table 7.1 presents the typical parameters of these layers; and Table 7.2 shows the typical parameters of transistors.

Table 7.1

Typical Parameters of Integrated *npn* Transistor Layers

Layer	N , cm^{-3}	d , μm	ρ , $\Omega \text{ cm}$	R_s , Ω/square
Substrate of p type	1.5×10^{15}	300	10	—
Buried n^+ layer	—	5-10	—	8-20
Collector n layer	10^{16}	10-15	0.5	500
Base p layer	5×10^{18}	3.0	—	200
Emitter n^+ layer	10^{21}	2	—	5-15

Note: N is the impurity concentration (surface concentration for diffused base and emitter layers), d is the layer depth, ρ is the material resistivity, and R_s is the sheet resistance.

The quantity R_s given in Table 7.1 is called a *sheet resistance*. The origin of this parameter is the following. Let us have a rectangular stripe of a material of length a , width b , and thickness d . If the current flows along the stripe, that is, parallel to its surface, the stripe resistance may be written in the form

$$R = \rho (a/bd) = R_s (a/b) \quad (7.1)$$

Table 7.2

Typical Parameters of Integrated *npn* Transistors

Parameter	Rating	Tolerance, δ %
Current gain B	100-200	± 30
Cutoff frequency f_T , MHz	200-500	± 20
Collector capacitance C_c , pF	0.3-0.5	± 10
Breakdown voltage V_{cb} , V	40-50	± 30
Breakdown voltage V_{eb} , V	7-8	± 5

where $R_s = \rho/d$. If the layer is inhomogeneous over its thickness (a diffused layer, for example), the quantity R_s will be written in the general form

$$R_s = \left[\int_0^d \sigma(x) dx \right]^{-1}$$

where $\sigma(x) = 1/\rho(x)$ is the material conductivity in the plane located at a distance x from the surface.

If $a = b$, the rectangular stripe takes the shape of a square, and its resistance becomes equal to R_s . Hence, the quantity R_s can be defined as a *lateral resistance of a layer or film in the shape of a square*. To stress the latter reservation, the actual dimension "Ohm" is replaced by "Ohm/ \square ", or Ohm/square. Knowing the quantity R_s , it is easy to calculate the resistance of a layer or film rectangular in shape from the given values of a and b .

Table 7.2 shows that the breakdown voltage of an emitter junction is one-fifth to one-seventh that of the collector junction. This feature, which is specific to all drift transistors, stems from the fact that the layers forming the emitter are of a lower resistance than the collector layers. With the transistor connected in a common emitter circuit, the breakdown voltage of the collector junction decreases according to Eq. (4.27). If the base is rather thin ($w < 1 \mu\text{m}$), a breakdown is generally caused by the punch-through effect, and the breakdown voltage is described by Eq. (4.28).

7.3.3. Parasitic parameters. Fig. 7.14a shows the simplified structure of an integrated *npn* transistor produced by the method of isolation diffusion. The integrated transistor structure consists of four layers, including the substrate: along with the working emitter and collector junctions, the transistor has a third (parasitic) junction between the collector *n* layer and the *p*-type substrate. The buried n^+

layer (not shown in Fig. 7.14a) does not make principal changes in the structure.

Since IC p -type substrate is *always connected to the most negative potential*, the voltage across the "collector-substrate" junction is always reverse or, in the worst case, close to zero. It is thus safe to replace this junction by the barrier capacitance $C_{c\ sub}$ shown in Fig. 7.14a.

The capacitance $C_{c\ sub}$ and the series collector resistance r_{sc} form an RC circuit connected to the active area of the collector. The equivalent circuit of the integrated nnp transistor now appears as shown in Fig. 7.14b.

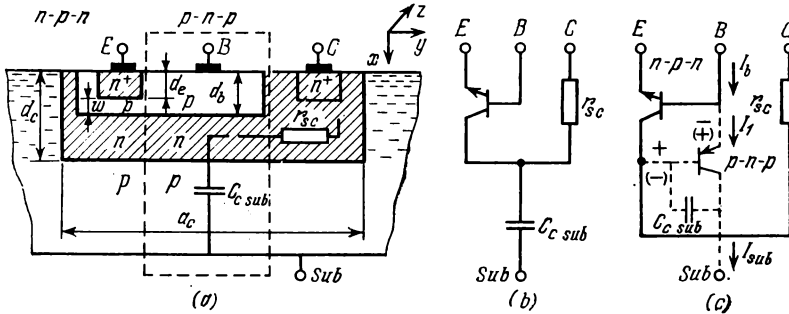


Fig. 7.14. The nnp transistor

(a) simplified structure showing parasitic pnp transistor; (b) simplified circuit model; (c) complete circuit model

The $r_{sc}C_{c\ sub}$ circuit that shunts the collector is the main feature of an integrated nnp transistor. This circuit naturally decreases the speed of response of the transistor and limits its cutoff frequency and switching time.

Since the substrate is at a constant potential, we can consider it to be grounded for ac components. So, inserting the $r_{sc}C_{c\ sub}$ circuit into the small-signal common-base circuit model (see Fig. 4.16) and disregarding the resistance r_b , we arrive at the conclusion: the capacitance $C_{c\ sub}$ adds to capacitance C_c and the resistance r_{sc} to the external resistance R_c (see p. 149). The equivalent time constant of (4.66) will then be written as

$$\tau_{\alpha oe} = \tau_{\alpha} + (C_c + C_{c\ sub})(r_{sc} + R_c) \quad (7.2)$$

From (7.2) it is clear that the parasitic parameters $C_{c\ sub}$ and r_{sc} set a limit on the speed of response of an integrated transistor under the ideal conditions at which $\tau_{\alpha} = 0$, $C_c = 0$, and $R_c = 0$.

In this case the equivalent time constant $\tau_{\alpha oe}$ is equal to the substrate time constant:

$$\tau_{sub} = C_{c\ sub}r_{sc} \quad (7.3)$$

Thus if $C_{c\ sub} = 2$ pF and $r_{sc} = 100\ \Omega$, then $\tau_{sub} = 0.2$ ns and the respective cutoff frequency $f_{sub} = 1/2\ \pi\tau_{sub} \approx 800$ MHz. If we consider the parameters τ_α and C_c and also include the external resistance R_c , the equivalent time constant will be greater and the cutoff frequency smaller.

The resistance r_{sc} taken equal to $100\ \Omega$ in the last example is a typical value for transistors without the buried n^+ layer. With the buried layer being present, the typical value of r_{sc} is $10\ \Omega$. In this case the time constant τ_{sub} decreases by a factor of ten and the effect of the substrate becomes of little significance.

The relation between $C_{c\ sub}$ and C_c primarily depends on the relation between the areas of respective junctions and the impurity concentration in the substrate and collector layers. Commonly, $C_{c\ sub} = 2$ or $3\ C_c$.

In calculating the capacitance $C_{c\ sub}$, one should take into account not only the bottom portion of the collector-substrate junction but also its side (vertical) portions (see Fig. 7.14a). The per-unit area capacitance of side portions is larger than that of the bottom because the acceptor concentration in isolating layers grows from the bottom of the junction to the surface (in Fig. 7.14a, the density of hatching reflects the variations in acceptor concentration). The typical value of per-unit area capacitance for the bottom is $C_{0x} = 100$ pF/mm², and for the side portions $C_{0y,z} = 150$ to 250 pF/mm². All the three components of the capacitance $C_{c\ sub}$ are commonly almost equal and lie in the range from 0.5 to 1.5 pF.

The passive base area combined with the underlying collector and substrate areas can be represented by a *parasitic npn* transistor. In Fig. 7.14a, the structure of such a transistor is surrounded by a dash line, and the equivalent circuit characterizing the interaction between the working *nnp* transistor and the parasitic *pnp* transistor is shown in Fig. 7.14c.

If the *nnp* transistor operates in the normal active region ($V_{cb} > 0$), the parasitic transistor stays cut off ($V_{cb} < 0$, see plus and minus signs unbracketed). In this mode of operation, the collector junction of the parasitic transistor is represented by the capacitance $C_{c\ sub}$ (see Fig. 7.14b). If the *nnp* transistor passes to the inverse region or the region of double injection ($V_{eb} < 0$), then the parasitic *pnp* transistor enters the active region ($V_{eb} > 0$, see the signs in brackets). The current $I_{sub} = \alpha_{pnp}I_1$ then flows into the substrate, where I_1 is a component of base current (see Fig. 7.14c).

The leakage of base current to the substrate impairs the parameters of the transistor operating in the double injection mode (see Sec. 8.2). For this reason transistors intended for work in this mode are specially doped with gold. The atoms of gold diffused into silicon play the role of traps; they decrease the lifetime of carriers. The value of α_{pnp} then drops below 0.1 , and I_{sub} becomes negligible.

The technique of dielectric isolation eliminates the parasitic *pnp* transistor, but the capacitance $C_{c\ sub}$ remains the same. As noted earlier this isolation technique ensures a smaller value of $C_{c\ sub}$ than *pn* junction isolation. If SiO_2 acts as a dielectric, the per-unit area capacitance is about 35 pF/mm^2 at a thickness of $1\text{ }\mu\text{m}$.

7.3.4. Typical fabrication procedure. Industry supplies the design engineer with polished and chemically treated wafers ready for use in the manufacturing process. We thus have a *p*-type silicon wafer whose polished surface is coated with a thin, natural layer of

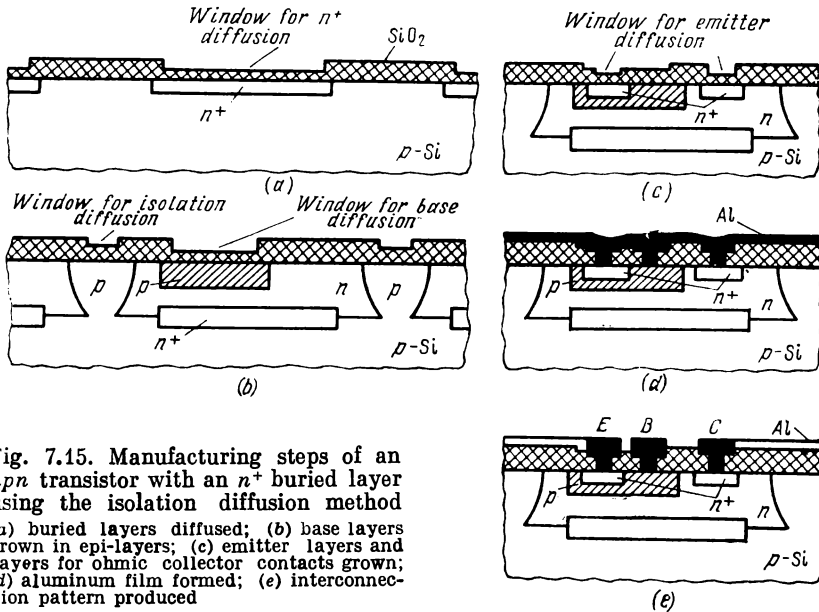


Fig. 7.15. Manufacturing steps of an *nnp* transistor with an n^+ buried layer using the isolation diffusion method (a) buried layers diffused; (b) base layers grown in epi-layers; (c) emitter layers and layers for ohmic collector contacts grown; (d) aluminum film formed; (e) interconnection pattern produced

oxide. Using the batch processing technique, we have to fabricate in this wafer transistors of the structure shown in Fig. 7.6b.

The sequence of steps will be as follows¹.

1. Total oxidation of the wafer.
2. First photomasking operation to define windows in the oxide for the diffusion of buried n^+ layers.
3. First diffusion to grow buried n^+ layers, as shown in Fig. 7.15a, using the diffusant arsenic or antimony.
4. Etching of the oxide from the entire surface.

¹ The sequence of stages given here does not include numerous operations involved in cleaning and rinsing of the wafer and in deposition and removal of the photoresist.

5. Growing of an n epitaxial layer, in which process the buried n^+ layer diffuses a little both into the substrate and into the epi-layer (see p. 217).

6. Total oxidation.

7. Second photomasking operation to open windows in the oxide for isolation diffusions.

8. Second diffusion to provide isolating p layers and isolated n islands in the epi-layer (see Fig. 7.5), using the diffusant boron.

9. Third photomasking operation to open windows in the oxide for base diffusion.

10. Third diffusion to grow base p layers (Fig. 7.15b), using the diffusant boron. The diffusion involves two stages, predeposition and drive-in (see p. 181).

11. Fourth photomasking operation to open windows in the oxide for emitter diffusion and ohmic contacts of the collectors.

12. Fourth diffusion to grow n^+ layers (Fig. 7.15c), using the diffusant phosphorus. The diffusion here is sometimes of the two-stage type too.

13. Fifth photomasking operation to define windows in the oxide for ohmic contacts.

14. Total deposition of aluminum film onto the wafer (Fig. 7.15d).

15. Sixth photomasking operation to open windows in the photoresist for the interconnection pattern.

16. Etching of the aluminum film through the photoresist mask and removal of the remaining photoresist (Fig. 7.15e).

17. Thermal treatment for firing of aluminum into silicon.

We omit here the assembly operations discussed earlier in Sec. 6.10. Let us make a few remarks on the fabrication procedure outlined above. Industry has recently begun to supply wafers with preliminarily grown epitaxial and buried n^+ layers. The first five operations in the procedure thus become superfluous.

We have noted in 10 that the boron diffusion for growing base layers includes two stages. This might appear to complicate the manufacturing process. But the approach is justifiable and conventional.

Indeed, in order that the injection efficiency of the emitter junction should be not below 0.999, the impurity concentration in the emitter layer must exceed not less than 100 times the concentration in the base layer [see Eqs. (4.22)]. Meanwhile, boron and phosphorus differ merely threefold in solid solubility at optimum temperatures (see Table 6.1). To overcome this contradiction, we must lower the surface concentration of boron. This can be done by one of the few methods.

We can diffuse boron at such a low temperature that its solubility can be a hundredth that of phosphorus. But then the diffusion coefficient will decrease by a few orders of magnitude, and so the dif-

fusion will take a few days or even a few weeks. We can also reduce the temperature in the diffusant source zone and thus create a "diffusant deficiency" near the wafer surface. But this process is difficult to control. So the two stage diffusion proves an optimum solution: during the impurity redistribution stage (the drive-in stage) the surface concentration is easy to decrease by a few orders of magnitude (see Fig. 6.7b).

The temperature of impurity redistribution is kept 150 to 200°C above the predeposition temperature in order to raise the impurity diffusion coefficient and cut down the time required for the process. The predeposition stage generally takes 20 to 40 min at 1 000 to 1 050°C, and the drive-in stage lasts a few hours at about 1 200°C.

Phosphorus introduction at the stage of emitter diffusion (see 12) is the last high-temperature operation in the fabrication procedure; the temperature is held 100 to 150°C below the temperature of boron distribution in order to keep the collector pn junction depth invariable. The diffusion time determines the n^+ layer thickness, and hence the transistor base width. In modern planar transistors, the typical base width is 0.5 to 0.7 μm .

Let us note in conclusion that as a result of repeated operations of photomasking, oxidation, and diffusion the surface of the oxide film before metallization becomes rather irregular. This often causes difficulty in ensuring good adhesion of aluminium to the oxide surface. In figures illustrating the structure (cross section) of transistors or ICs, the relief of the oxide film is not shown for simplicity.

7.4. Varieties of NPN Transistors

The progress of microelectronics has brought about the evolution of some varieties of $n pn$ transistors. There are no analogs of these transistors in discrete transistor engineering. Consider the most important of these varieties.

7.4.1. Multiemitter transistor. The structure of a multiemitter transistor (MET) appears in Fig. 7.16a. These transistors form the basis of a rather widespread class of digital integrated circuits called transistor-transistor logic (TTL) circuits, which will be described later in Subsec. 10.2.5. This type of transistor can have 5 to 8 emitters and even more.

To a first approximation, the multiemitter transistor can be regarded as a combination of individual transistors with interconnected bases and collectors (Fig. 7.16b). The features of the multiemitter transistor as a single structure are as follows (Fig. 7.16c).

First, each pair of adjacent emitters together with the base p layer that isolates these emitters forms a *lateral n^+pn^+* transistor. Let one of the emitters be **forward** biased and the adjacent one **reverse**

biased. The first will then inject electrons and the second collect those injected electrons which have passed through the side surface of the first and reached the second without recombination. Such a *transistor effect* is **parasitic** for the multiemitter transistor: a current will flow through the reverse-biased junction, which must be in the off condition. To avoid the lateral transistor effect, the distance between the emitters must generally exceed the diffusion length of carriers in the base layer. If the transistor is doped with

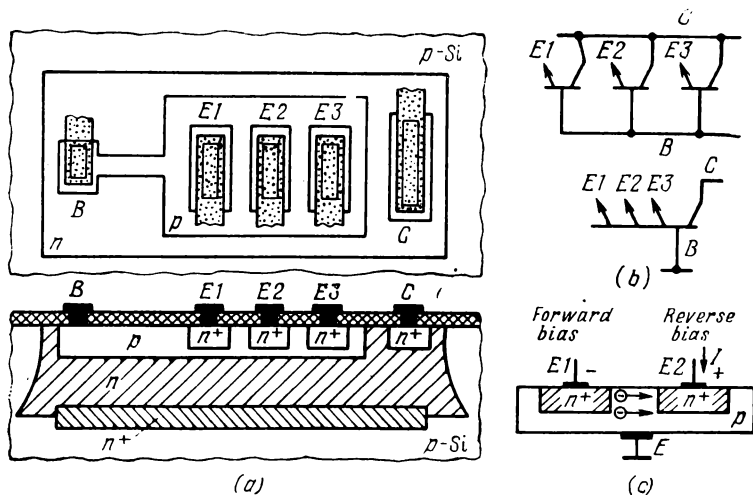


Fig. 7.16. Multiemitter transistor

(a) layout and structure; (b) symbols; (c) interaction between adjacent emitters

gold (see p. 226), the diffusion length does not exceed 2 or 3 μm , and so a distance of 10 to 15 μm proves practically acceptable.

Second, it is important that the multiemitter transistor should have as small an inverse current gain as possible. Otherwise, with the emitters reverse biased and the collector forward biased, a significant amount of the carriers injected by the collector will reach the emitters. A current that will flow in the emitter circuits can produce a parasitic effect similar to that described above.

As known, the inverse current gain is always smaller than the normal one because of the difference in the doping levels and in the areas of the emitter and collector (see p. 130). To decrease the inverse factor α_1 of the multiemitter transistor still more, the resistance of the passive base is intentionally increased by placing the ohmic base contact further away from the active area of the base (see Fig. 7.16a). With such a transistor geometry, the resistance of the narrow "neck" between the active area and base contact can

reach 200 to 300 Ω and the voltage drop across this neck resulting from the base current flow can be 0.1 to 0.15 V. So the forward voltage on the collector junction (operating in the inverse region) will be 0.1 to 0.15 V smaller in the active area than that near the base contact. Correspondingly, the injection of electrons from the collector into the active base area becomes insignificant and the parasitic emitter currents disappear.

7.4.2. Multicollector *n*pn transistors. The structure of a multicollector transistor (MCT) shown in Fig. 7.17a does not differ from the structure of a multiemitter transistor. The difference lies in the

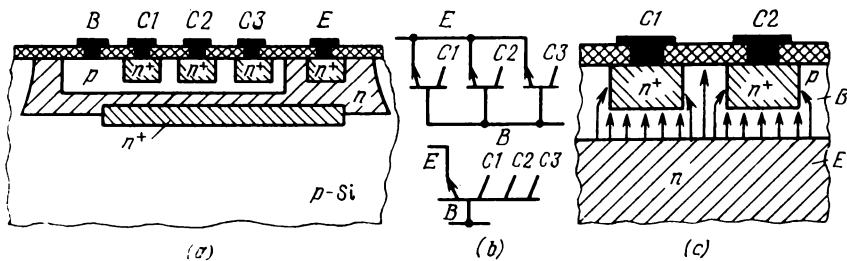


Fig. 7.17. Multicollector transistor

(a) structure; (b) equivalent circuits; (c) trajectories of injected carriers

method of using the structure. It is safe to say that *the MCT is the MET operated in the inverse region*: the epitaxial *n* layer serves as a common emitter, and heavily doped small-area *n*⁺ layers act as collectors. Such a layout forms the basis of one of the most popular classes of digital integrated circuits called integrated injection logic (I²L) circuits discussed in Sec. 10.3¹. The equivalent circuit of a multicollector transistor is given in Fig. 7.17b.

The main problem involved in the development of a multi-collector transistor is to increase its normal current gain from the common n emitter to each of the n⁺ collectors. Naturally, this problem is the reverse of that which related to the multiemitter transistor, where we attempted to reduce the current gain from the n layer to the n⁺ layers.

For this type of transistor, it is desirable that the buried *n*⁺ layer can lie as close to the base as possible, or simply be in contact with the base, which is the case when using CDI isolation technique. The heavily doped *n*⁺ layer acting as an emitter will then secure a high injection efficiency. As regards the increase of the transport

¹ Multicollector transistors employed in I²L circuits have the same basic properties as the transistors considered below and only differ in the method of power supply.

factor, the n^+ collectors should be spaced at the smallest distance possible to reduce the area of the passive base region. Both these approaches certainly have limitations due to design and manufacturing factors. Nevertheless, even if the space between the collectors is comparatively large, there is a possibility for ensuring α of 0.8 or 0.9 or B of 4 to 10, both being related to the entire set of collectors. This is sufficient for proper functioning of I²L circuits if the transistor has not more than 3 to 5 collectors. The current gain per collector is equal to the total gain divided by the number of collectors. Thus for the given values of total current gain B , the gain per collector is in excess of unity, which is enough for I²L functioning.

Figure 7.17c shows the trajectories of injected carriers in the base. As seen, the fraction of carriers getting onto the collectors is significantly larger than would be found if we calculate this share from the formal collector-to-emitter area ratio. This explains why the real current gain B has comparatively high values, as given above. In calculating the gains α and B , therefore, one should use the **effective** rather than the geometric areas. We have mentioned this fact on p. 130 in discussing a transistor operated in the inverse mode.

From Fig. 7.17c it is also seen that the mean carrier trajectory length noticeably exceeds the active base thickness w . That is why the mean diffusion time will be substantially smaller than for multiemitter and individual transistors having the same value of w [see Eq. (4.45)]. The difference in transit time is still greater because *the multicollector transistor has a retarding rather than an accelerating base field for injected carriers*. The transit time t_{tr} is not less than 5 to 10 ns, and the respective cutoff frequency f_T does not exceed 20 to 50 MHz (compare with the parameters in Table 7.2).

On the other hand, the collector capacitance C_c in multi-collector transistors is noticeably lower than that in multiemitter and conventional transistors because of the small area of an n^+ collector. Therefore, the terms $C_c R_c$ and $C^* R_c$ in Eqs. (4.66) and (4.67) can often be neglected.

7.4.3. Schottky-barrier transistor. The purpose and the principle of this transistor are discussed in Sec. 8.5. The structure of an integrated Schottky-barrier transistor is shown in Fig. 7.18. The layout here gives elegant solution to the problem of combining the transistor with the Schottky diode: the aluminum interconnection pattern that ensures ohmic contact with the base p layer extends toward the collector n layer. At first glance, the aluminum stripe shorts out the collector and the base layer. But in reality the stripe forms a **nonrectifying**, ohmic contact with the base p layer, and a **rectifying**, Schottky contact with the collector n layer (see Sec. 3.3). That is why the equivalent circuit of this structure is similar to the circuit model of Fig. 8.12.

The structural arrangement of Fig. 7.18 is certainly applicable not only to simplest transistors but also to multiemitter transistors.

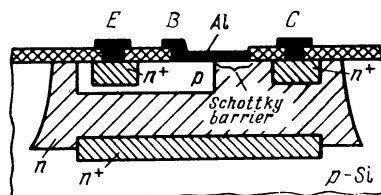


Fig. 7.18. Schottky-barrier transistor

In either type, the storage and removal of excess carriers are excluded, which decreases the time of switching from the completely turn-on to the cut-off state by a factor of 1.5 to 2.

7.4.4. Superbeta transistor. The superbeta transistor, as the name implies, features a very high common-emitter current gain B that ranges from 3 000 to 5 000 and over owing to a superthin base width w , equal to 0.2 or 0.3 μm .

The fabrication of a superthin base presents a serious technological problem. First, the base width is the difference in depth between

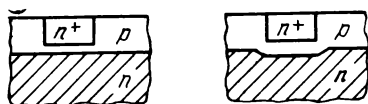


Fig. 7.19. Illustrating the manufacture of a superthin base

the base and emitter layers: $w = d_b - d_e$. If the tolerance on the base width is $\pm 10\%$, or 0.02 μm , then with a base layer depth d_b of 2 μm , the emitter base depth d_e must be equal to $1.8 \pm 0.02 \mu\text{m}$. So the diffusion process must produce an emitter layer to an accuracy of 1.25%, which is hardly within the technological possibilities. Second, in the course of emitter diffusion, when the metallurgical boundary of the emitter approaches that of the collector to within 0.4 μm , another difficulty arises: a further diffusion of phosphorus atoms in the emitter layer is accompanied by the diffusion (at the same rate) of boron atoms in the base layer. The emitter layer may be said to "drive down" the metallurgical boundary of the previously grown base layer (Fig. 7.19), thereby preventing the base from getting thinner than 0.4 μm . It took engineers many years to overcome the above difficulties and ensure the **reproducibility** of superthin base width.

A high current gain of superbeta transistors is achieved at a sacrifice in the breakdown voltage, which is merely 1.5 to 2 V. A low breakdown voltage is the result of junction punch-through inherent in transistors

with a thin base (see Subsec. 4.4.5). *Superbeta transistors are thus not universal but special elements of an integrated circuit.* The main field of application of these transistors is in the input stages of operational amplifiers (see Sec. 10.10).

It should be noted that a still further decrease in the base width to $0.2\text{ }\mu\text{m}$ and below is more dependent on the physical rather than on the manufacturing problems. To illustrate the point, let us assume that the average acceptor concentration in the base is equal to $8 \times 10^{15}\text{ cm}^{-3}$ (see Fig. 7.12). The acceptors will then number 2×10^5 per centimeter of length. With a base width of $0.1\text{ }\mu\text{m}$, or 10^{-5} cm , the base will consist of merely two layers of acceptor atoms. In this case the notion of impurity concentration gradient in the base layer and the related notion of built-in field becomes meaningless. Besides, the processes of carrier scattering and the laws of carrier motion in the base change qualitatively. The classical theory of transistors thus becomes largely invalid.

7.5. PNP Transistors

The fabrication of *pnp* transistors whose parameters would match the *npn* transistor parameters is a problem that still remains to be solved. That is why the available range of integrated *pnp* transistors is much inferior to that of *npn* transistors in current gain and cutoff frequency.

Other conditions being equal, *pnp* transistors are known to have approximately one-third the cutoff frequency of *npn* transistors

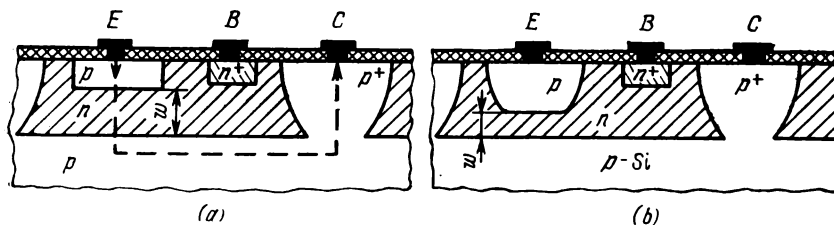


Fig. 7.20. Parasitic *pnp* transistors

(a) emitter formed at the stage of base diffusion for *npn* transistor; (b) emitter specially grown by deep boron diffusion

because of a lower hole mobility as against the electron mobility. But modern technology of *pnp* transistors does not as yet permit approaching those "equal conditions" which would ensure just a threefold difference in cutoff frequency.

At an early stage of IC development, *pnp* structures formed of the layers of a base, collector, and substrate served as *pnp* transistors (Fig. 7.20a). These transistors are generally called *parasitic* by analo-

gy to the transistors which enter into the structure of IC *nnp* transistors (see Fig. 7.14a). There is no need for additional operations in producing a parasitic *pnp* transistor, but its parameters prove extremely poor because of a large base width (comparable with the epi-layer thickness) and a low level of emitter doping. With a base width of 10 μm , the cutoff frequency f_T is 1 or 2 MHz, and the gain B is 2 or 3.

An equally severe drawback of parasitic *pnp* transistors is that the isolating p^+ layer is linked with the substrate and with other isolating layers via the substrate. Consequently, *the collectors of all the pnp transistors become coupled together*. This greatly limits the field of application of *pnp* transistors (compare Fig. 7.20 with Fig. 7.4).

An added manufacturing operation—**deep** acceptor diffusion (Fig. 7.20b)—makes it possible to obtain a smaller base width and raise the value of B to 8–10 and f_T to 4–6 MHz. But this approach lengthens the period of diffusion and fails to rectify the drawback of collectors coupling via the substrate.

At present, the basic type of *pnp* transistor structure is a **lateral pnp** transistor (Fig. 7.21). This transistor is isolated from other elements, has much better parameters than the parasitic *pnp*

transistor, and its technology is quite congruous to the conventional fabrication procedure that uses isolation diffusion.

Emitter and collector layers are grown at the stage of base diffusion. The collector encircles the emitter and thus collects the holes injected from all the **side** portions of the emitter layer. The surface side portions of p layers have an increased impurity concentration, which raises the injection efficiency. Since the base diffusion gives rather **shallow** layers, 2 or 3 μm thick, the base width (the distance between p layers) can be 3 or 4 μm . Consequently, the cutoff frequency can reach 20 to 40 MHz, and the current gain B up to 50.

From Fig. 7.21 it is clear that the lateral *pnp* transistor, like the parasitic transistor, is not of the **drift** but of the **diffusion type** because its base (the epitaxial n layer) is **homogeneous**. This factor along with a lower hole mobility accounts for the fact that the frequency and transient response of the *pnp* transistor is approximately

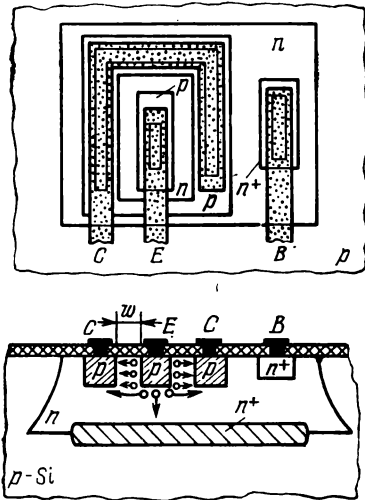


Fig. 7.21. Topology and structure of a *pnp* lateral transistor

a factor of 10 below that of the drift transistor even at the same base width. It is also apparent from Fig. 7.21 that in order to increase the current gain α , the bottom area of the emitter layer should be small in comparison with the area of its side parts. So the emitter layer need be made as narrow as possible; the width of the window for diffusion of this layer is 3 to 5 μm .

Note that the structure of a lateral *pnp* transistor shows an **electro-physical** symmetry because the emitter and collector layers are of the same type. This means, in particular, that the breakdown voltages of emitter and collector junctions are equal and commonly range from 30 to 50 V. The normal and inverse current gains, β_N and β_I , are also close in value.

The lateral structure of a *pnp* transistor allows for an easy conversion of this transistor into a **multicollector** *pnp* transistor. For

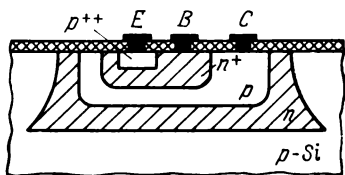


Fig. 7.22. Vertical *pnp* transistor

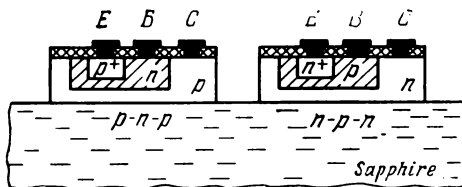


Fig. 7.23. A *pnp* transistor fabricated compatibly with an *npn* transistor by the SOS technique

this it is enough to divide the circular p collector (see Fig. 7.21) into m portions and provide each with separate terminations.

The current gain of each collector will be approximately a factor of m smaller than for the entire circular collector, but all the collectors will operate independently, and their loads will be isolated, that is, "decoupled" from each other.

Major disadvantages of the lateral *pnp* transistor are a comparatively large base width and its homogeneity. These disadvantages can be avoided in a **vertical** structure (Fig. 7.22) but at the cost of extra operations added to the fabrication process.

As seen from the structure of Fig. 7.22, two such operations are added: deep diffusion of a p layer and final diffusion of a p^{++} layer. The last operation is rather problematic because the p^{++} layer should be formed from an acceptor material whose solubility must be higher than that of phosphorus used for the deposition of the n^+ layer. Since such acceptor materials are practically nonexistent, it is necessary to etch off the upper, most heavily doped portion of the n^+ layer before depositing the p^{++} layer, which makes the manufacturing process more complex.

The silicon-on-sapphire (SOS) technology offers great potential in the fabrication of high-quality *pnp* transistors (see Subsec. 7.2.3). The SOS technology (Fig. 7.23) can produce a *pnp* transistor essentially **independent** of an *nnp* transistor, starting from the epitaxial growth of the *p* layer (the *n* and *p* layers are grown **selectively** through **different** masks). This approach permits optimizing both the base width and the level of doping of the emitter layer. But since there is a need for local epitaxial growth and two additional diffusions, the fabrication procedure is rather complex and costly.

7.6. Integrated Diodes

Any of the two *pn* junctions, either an emitter or collector junction, located in the isolating islands can act as a diode. It is also possible

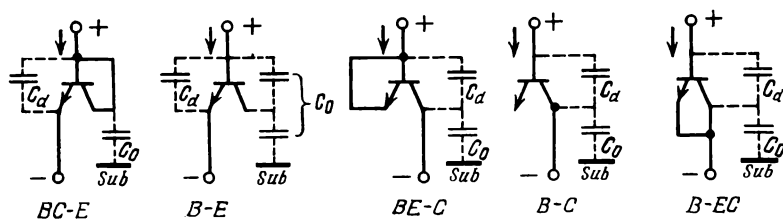


Fig. 7.24. Integrated diodes, or transistors suitably connected to perform a diode action

to use combinations of the junctions. The *integrated diode is thus essentially an integrated transistor connected in such a configuration as to perform a diode action.*

There are five possible ways of connecting a transistor to form a monolithic diode (Fig. 7.24). Table 7.3 gives the typical param-

Table 7.3

Typical Parameters of Integrated Diodes

Parameter	Diode type				
	BC-E	B-E	BE-C	B-C	B-EC
V_{br} , V	7-8	7-8	40-50	40-50	7-8
I_{rev} , nA	0.5-1	0.5-1	15-30	15-30	20-40
C_d , pF	0.5	0.5	0.7	0.7	1.2
C_0 , pF	3	1.2	3	3	3
t_r , ns	10	50	50	75	100

ters for each of the five configurations. The designations of diodes are as follows: the letter ahead of the hyphen denotes an anode, and the letter after the hyphen a cathode; the two letters identifying two layers tied together are not hyphenated. The data of Table 7.3 show that five diode variants differ both in static (dc) and dynamic (ac) parameters.

Breakdown voltage V_{br} depends on the junction used. As seen from Table 7.2, the diodes using an emitter junction show a lower breakdown voltage.

Reverse currents I_{rev} (disregarding leakage currents) are thermally generated currents in the junctions. They depend on the **volume** of a junction and thus are lower in those diodes which use **only** an emitter junction having the smallest area.

Diode capacitance C_d (capacitance between the anode and cathode) depends on the area of junctions used; this capacitance is maximum in a diode having its junctions connected in parallel (B-EC diode). The *parasitic capacitance on the substrate*, C_0 , shunts the anode or cathode to ground (the substrate being regarded as grounded). The capacitance C_0 generally coincides with the capacitance $C_{c\ sub}$ we have dealt with in discussing the *npn* transistor (see Fig. 7.14b). In the B-E diode, however, the capacitances $C_{c\ sub}$ and C_c are found to be connected in series, and so the resultant capacitance C_0 is a minimum.

Recovery time t_r of reverse current (the time a diode requires to go from the on to the off condition) is minimum in a BC-E diode; this diode stores the charge **only** in the base layer because the collector junction is short-circuited. In other types, the charge is stored not only in the base but also in the collector, for which reason the time taken for charge removal is longer.

The comparison of individual diodes allows us to come to the conclusion that the BC-E and B-E types are on the whole *most optimal*. Low breakdown voltages of these diodes are of little significance in low-voltage ICs. The BC-E diode is the most common type.

Apart from diodes proper, *reference diodes* find widespread application in ICs. They also come in a few versions, each designed for a definite stabilizing voltage and temperature coefficient.

Where it is necessary to maintain voltage at 5 to 10 V, use can be made of a B-E diode to operate under **reverse bias** in the breakdown region; the temperature sensitivity of the device ranges from 2 to 5 mV° C⁻¹. Where voltages should be kept at 3 to 5 V, it is possible to employ either a **reverse-biased** BE-C diode designed for operation in the punch-through region (see p. 133) or a **reverse-biased pn** junction specially formed in an isolating layer (Fig. 7.25a). In the latter device, the n^+ layer is produced at the stage of emitter diffusion. Since the surface portion of the insulating layer is strongly

doped, the junction has a p^+n^+ structure and displays a tunnel breakdown (**low-voltage** breakdown). The temperature sensitivity ranges from -2 to -3 mV $^{\circ}\text{C}^{-1}$.

Very popular now are reference diodes rated at voltages which are equal to or a multiple of the forward voltage on the junction,

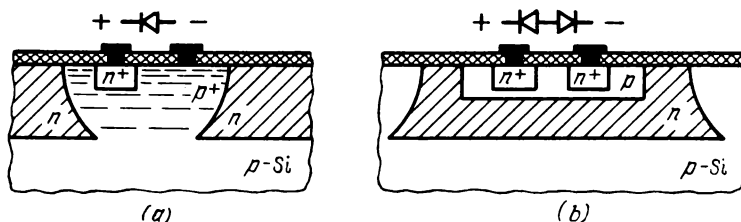


Fig. 7.25. Integrated stabilizer diodes

(a) in an isolating layer; (b) in a base layer, using temperature compensation

$V^* \approx 0.7$ V. These types can use one or a few series-connected BC-E diodes operating in the **forward-bias** conditions. The temperature sensitivity is -1.5 to -2 mV $^{\circ}\text{C}^{-1}$.

The structure of Fig. 7.25b illustrates two pn junctions grown in the base layer. On applying a voltage across the n^+ layers, one junction will operate in the avalanche region and the other in the forward bias region. Such a structure is attractive for its low temperature sensitivity (± 1 mV $^{\circ}\text{C}^{-1}$). This is because the temperature sensitivities in avalanche breakdown and in the forward-bias condition are different in sign.

7.7. Junction Field-Effect Transistors

The field-effect transistors (junction FETs) discussed in Sec. 5.3 fit well the requirements of general bipolar IC technology, and therefore they are often fabricated simultaneously with bipolar transistors on one and the same chip. The typical structures of FETs formed in isolated islands are shown in Fig. 7.26.

In the structure of Fig. 7.26a, the p layer of the gate is grown at the stage of base diffusion, and the n^+ layers that provide an ohmic contact to the source and drain regions are formed at the stage of emitter diffusion. Note that the gate p layer completely encircles the drain, so that the current between the source and drain can flow only through the controllable channel.

In n islands intended for FETs, a buried p^+ layer is formed instead of the buried n^+ layer. This layer serves to reduce the initial channel thickness a and thus decrease the cut-off voltage [see Eq. (5.29)]. The formation of the buried p^+ layer necessitates additional manufacturing operations. In order that the buried p^+ layer might diffuse

For the source and drain regions to be connected **only** via the channel, the n^+ layer is made wider (in plan) than the p layer (see Fig. 7.26*b*). The n^+ layer then comes in contact with the n epitaxial layer and both form the “upper” and “lower” gates. In the lower part of Fig. 7.26*b* the “upper-to-lower” gate contact is shown by a dash line. The p substrate is connected to the most negative voltage.

7.8. MOS Transistors

The fabrication of MOS and bipolar transistors on the same silicon chip is generally possible, but this manufacturing process represents a special case. Bipolar and MOS ICs are generally developed and manufactured separately. These two types of IC serve to handle either different functions or one and the same function, but taking advantage of the appropriate class of transistors.

7.8.1. Simple MOS transistor. Since integrated MOS transistors do not require isolation (see Sec. 7.2), their structure does not differ externally from the structure of discrete counterparts. Fig. 7.27

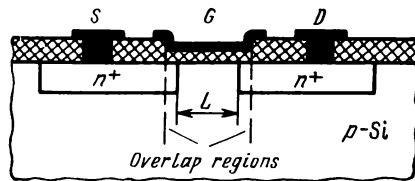


Fig. 7.27. MOS transistor with gate overlap

shows an induced n -channel MOS transistor structure analyzed in detail in Sec. 5.2. Consider some features of the MOS transistor as an integrated element.

From the comparison of Figs. 7.6, 7.21, and 7.22 it is primarily evident that the MOS transistor is easier to manufacture than the bipolar transistor. The fabrication procedure includes only one stage of diffusion and four photomasking processes to define windows for the diffusion, thin oxide, ohmic contacts, and interconnections. The simple technology ensures a higher yield and lower cost.

Since the structure is free from isolating islands, there is a possibility for closer spacing of the elements on the chip, thereby increasing the packing density. In the absence of isolation, however, the substrate becomes a common electrode for all transistors. This situation may result in different parameters of externally identical transistors. Indeed, if the substrate is at a constant potential, whereas the sources of transistors have different potentials (such a difference is typical of many circuits), the voltages V_{sub} between the substrate and source will also be different. According to Eq. (5.15),

this is equivalent to the difference between the threshold voltages of MOS transistors.

As known, the main factor that limits the speed of response of MOS transistors is commonly parasitic capacitances (see p. 166). The metallic interconnection layout employed in ICs is much denser than the wire-bonding layout specific to units and blocks based on discrete components. That is why the parasitic capacitances in an integrated MOS transistor are lower than for its discrete counterpart, and therefore the transient response of the former is a few times higher than that of the latter.

The parasitic capacitances of a MOS transistor were shown in Fig. 5.10. The barrier capacitances of source and drain junctions, $C_{sub s}$ and $C_{sub d}$, are calculated from Eqs. (3.34); with n^+ layers measuring $20 \times 40 \mu\text{m}$, these capacitances lie between 0.04 and 0.10 pF.

The specific capacitance of metallizations is determined from the elementary formula

$$C_{0m} = \epsilon_0 \epsilon / d \quad (7.4)$$

where d is the thickness of a protective oxide; and ϵ is the oxide permittivity. Substituting $\epsilon = 3.5$ and $d = 0.7 \mu\text{m}$, we find that a typical value of the capacitance is $C_{0m} \approx 50 \text{ pF/mm}^2$. With a stripe width of $15 \mu\text{m}$, the linear capacitance is 0.75 pF/mm . As seen, the stripes merely 50 to $100 \mu\text{m}$ in length can have a capacitance of 0.04 to 0.08 pF, which is comparable with junction capacitances $C_{sub s}$ and $C_{sub d}$. Bonding pads contribute still more to the total capacitance of the integrated MOS structure: a 100 by $100 \mu\text{m}^2$ pad has a capacitance of about 0.5 pF.

Overlap capacitances C_{gs} and C_{gd} (see Fig. 5.11) are not amenable to accurate calculation because the edges of the gate metallization and diffusion layer boundaries are irregular, and so the overlap area spread is significant (see Fig. 7.27). By setting the thickness d of a **thin** oxide layer equal to $0.12 \mu\text{m}$, we can estimate the order of magnitude of these capacitances. Using Eq. (7.4), we find that the per-unit area capacitance of thin oxide is $C_0 \approx 300 \text{ pF/mm}^2$. For a width of the source and drain taken equal to $50 \mu\text{m}$ and an overlap to $2 \mu\text{m}$, the capacitances $C_{gs} = C_{gd} \approx 0.03 \text{ pF}$. These values are smaller than the values of other capacitances, and therefore the capacitance C_{gs} may often be neglected. The capacitance C_{gd} , however, which is a *feedback capacitance* between the transistor output (drain) and input (gate), shows itself in many circuits as an **equivalent capacitance** KC_{gd} of a much higher value on account of the *Miller effect* (see Subsec. 9.5.4). The multiplier K is the amplification factor of a circuit, which can range from a few units to a few tens and more. For this reason the equivalent feedback capacitance

KC_{gd} can reach 0.1 to 0.5 pF, thus exceeding all other capacitances in value.

In **complementary** MOS transistor integrated circuits (CMOSICs), both n -channel and p -channel MOSFETs should be produced on the same wafer. Then one type of transistor need to be formed in a special isolating island. For example, if the silicon substrate is p -type, then an n -channel transistor can be obtained directly in the substrate, and a p -channel transistor in an n island (Fig. 7.28a). This island is in principle easy to realize, though the process involves extra

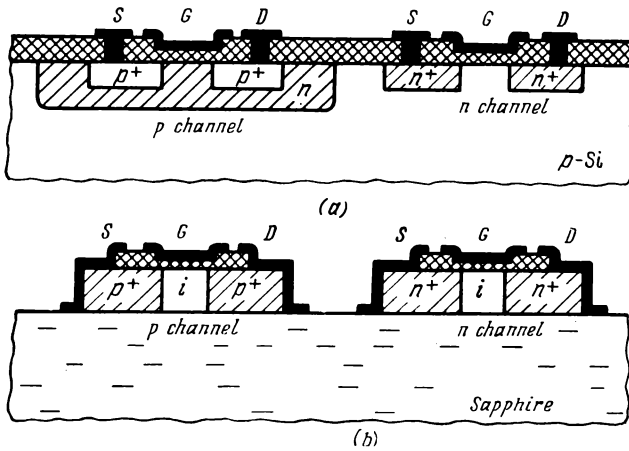


Fig. 7.28. CMOS transistors

(a) using n -type isolating island; (b) using air isolation by the SOS technique

operations such as photomasking, donor diffusion, and other manufacturing steps. Besides, it becomes difficult to form low-resistance p^+ layers in the upper, heavily doped, portion of the n island. Another way of fabrication of CMOS transistors on the common wafer is the SOS technique discussed in Subsec. 7.2.3. The islands of **intrinsic** silicon are first produced on the sapphire wafer. Next a donor impurity is diffused into some islands to produce n -channel transistors and an acceptor impurity into other islands to produce p -channel transistors (Fig. 7.28b). Though the number of fabrication steps here is greater than in the manufacture of transistors of one type only, the SOS technique obviates difficulties involved in the growth of low-resistance source and drain regions (see above).

As for the fabrication of MOS transistors and bipolar transistors compatibly on one wafer, the procedure is in principle simple (Fig. 7.29); n -channel transistors are made directly in the p substrate

at the stage of emitter diffusion, and p -channel transistors are produced in insulating islands at the stage of base diffusion.

In the course of development of microelectronics, the advancement in the performance of MOS transistors continued in two main directions: the first was toward an increase in the speed of response, and the second toward a decrease in the threshold voltage. The aim of the second trend was to reduce the working voltages of MOS transistors along with the dissipated power. Because the total power at the chip has an upper limit, a decrease in the power dissipated in

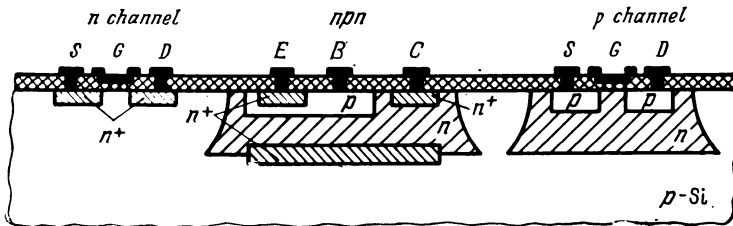


Fig. 7.29. Bipolar and MOS transistors made on the same chip

each transistor allows an increase in the level of integration, while a decrease in supply voltages facilitates the joint operation of MOS transistor ICs and low-voltage bipolar transistor ICs, thereby eliminating the need to use special matching circuits.

7.8.2. Ways of increasing the speed of response. Increasing the speed of response of MOS transistors involves first of all a decrease in overlap capacitances. The *technology of self-aligning gates* decreases overlap capacitance by approximately a factor of 10. The general idea of this technology lies in that the steps of manufacture of the source and drain regions follow rather than precede the step of fabrication of the gate. In this process, the gate acts as a mask for growing source and drain layers, so the edges of the gate coincide with the edges of these layers and the overlap does not appear.

One of the types of MOS transistor with a self-aligned gate is shown in Fig. 7.30. The sequence of manufacturing steps here is as follows. The first step involves the diffusion of n^+ layers intentionally spaced at a distance that exceeds the desired channel length. Oxidation then follows in the region between the n^+ layers and partially above them to produce a thin oxide. Next an aluminum gate electrode smaller in width than the gap between the n^+ layers is deposited onto the oxide layer. The last step includes ion doping (implantation of phosphorus atoms) through a mask formed by the aluminum gate and thick protective oxide. Phosphorus atoms penetrate the silicon through the thin oxide and "extend" the n^+ layers up to

the edge of the aluminum finger so that the gate edges practically coincide with the source and drain edges. The implanted layers are doped a little weaker than the n^+ diffusion layers, and therefore

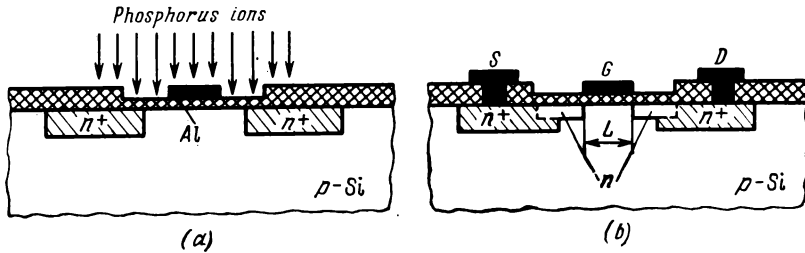


Fig. 7.30. MOS transistor with a self-aligned gate produced by ion implantation

they are designated as n layers. The implantation depth is also somewhat smaller than the diffusion depth, and is equal to 0.1 or 0.2 μm .

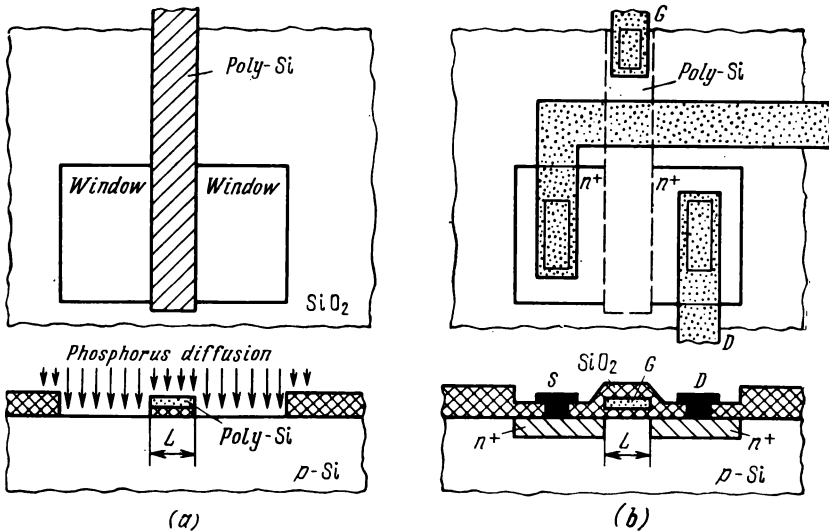


Fig. 7.31. Self-aligned poly-Si gate MOS transistor

(a) donor diffusion through mask having poly-Si layer; (b) finished structure after glassover (protective oxide layer deposition) and metallization

Another version of the MOS transistor with a self-aligned gate appears in Fig. 7.31. The first fabrication step necessitates etching of the oxide to define a window for the **entire structure** of the transistor. Silicon is then oxidized in the middle portion of the window

to produce a thin oxide stripe whose width should be equal to the desired length L of the channel. A poly-Si layer is now deposited on the oxide stripe. This layer is of the same width as the oxide stripe, but longer to enable it to extend beyond the edges of the original window in the oxide (Fig. 7.31a). The silicon being deposited has a rather low resistivity, so the polysilicon layer acts as a metallic gate in conventional MOS transistors. The next step is the shallow diffusion of a donor impurity through a mask formed by the poly-Si gate stripe and the protective oxide surrounding the window. This results in n^+ layers of the source and drain, whose edges almost coincide with the edges of the polysilicon gate. Further, the whole of the chip surface is oxidized and the oxide layer is etched as usual

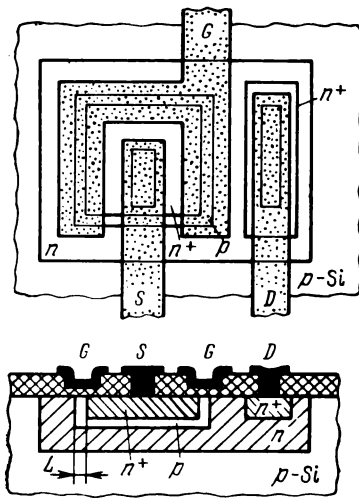


Fig. 7.32. MOS transistor produced by double diffusion

to open windows for ohmic contacts along with the contact for the silicon gate. The metallization step comes last to produce the desired interconnection pattern. As clear from Fig. 7.31b, the poly-Si gate lies "buried" in the protective oxide, the ohmic contact to the gate being positioned beyond the working region of the transistor.

A decrease in the parasitic capacitances of a MOS transistor, primarily overlap capacitance C_{gd} , brings to the forefront the problem of reducing the time constant of transconductance, τ_s . At low capacitances, τ_s can become the main factor that limits the response characteristics of transistors.

Conversion from p -channel to n -channel transistors permitted τ_s

to be reduced by approximately a factor of 3 owing to the increased carrier mobility. A further decrease in τ_s calls for a reduction in channel length L . The double diffusion process adequately fills the need. Fig. 7.32 shows the structure of a MOS transistor made by double diffusion.

This structure is similar to the structure of a conventional npn transistor (see Fig. 7.14a), the only, but rather substantial, difference being that the emitter n^+ layer (the source layer in the given case) has almost the same area as the base p layer (the channel layer here). To ensure the exact "driving" of the n^+ layer into the p layer, the diffusion of donors for the n^+ layer is made through the same window in the oxide as that which has been used for the diffusion

of acceptors to produce the p layer. This eliminates the necessity for photomask alignment in the two successive photomasking processes, and hence does away with the alignment error that might lead to the shift of the n^+ layer with respect to the n layer. So the distance between the n^+ and n layers (the p layer thickness) can approximately be the same as the base width w in an npn transistor, about $1\text{ }\mu\text{m}$ and less (see Fig. 7.14a).

Near the surface, the distance between n^+ and n layers plays the role of a channel length (see Fig. 7.32). In transistors having L equal to or smaller than $1\text{ }\mu\text{m}$ (as against 4 or $5\text{ }\mu\text{m}$ in most advanced MOSFETs made by traditional technology), the time constant τ_s according to Eq. (5.27) can be lower than 0.005 ns , and the cutoff frequency f_s higher than 30 GHz .

7.8.3. Ways of decreasing the threshold voltage. Transistors of the structure shown in Fig. 7.31 are generally referred to as MOS transistors with a *silicon gate*. They feature not only a small overlap capacitance but also a decreased threshold voltage, 1 or 2 V instead of 2.5 to 3.5 V as is commonly the case. This is because the material of the gate and that of the substrate are the same, namely, silicon. So the contact potential difference φ_{MS} between the gate and substrate is equal to zero; this leads to a decrease in the threshold voltage according to Eq. (5.3a). The use of a **molybdenum gate** gives about the same result.

Apart from the contact potential difference, other parameters entering into Eqs. (5.3) can be varied to decrease the threshold voltage. Thus it is possible to replace a thin SiO_2 layer by a thin **sputtered** layer of silicon nitride, Si_3N_4 , whose permittivity is about twice as large as that of SiO_2 (nearly 7 against 3.5). This tends to increase the per-unit area capacitance C_0 and thus reduce the corresponding components of the threshold voltage. Silicon nitride, serving as a gate insulator, also offers other advantages, such as lower noise, a higher time stability of the I - V characteristic, and enhanced radiation stability of the MOS transistor.

It is also possible to use (100) Si substrates instead of the traditional (111) substrates. This raises the density of surface states (see Fig. 2.5) and hence increases the charge of electrons captured by these states. The negative addend Q_{os}/C_0 in Eq. (5.3a) consequently grows and the algebraic sum of both terms, namely, the voltage V_{0F} , decreases.

Acceptor atoms, when introduced into the thin oxide, have a reverse effect. They capture from the silicon surface a fraction of electrons generated by **donor** impurities, which are always present in the oxide (see p. 113). The negative charge Q_{os} thus drops off. Acceptors can be introduced into the oxide by ion implantation.

The combination of the above methods allows a decrease in threshold voltage to arbitrarily small values.

It should be borne in mind, however, that too low values of threshold voltage, down to 0.5 to 1 V and below are in most cases unacceptable for circuit engineering reasons (because of low noise immunity.)

7.8.4. MNOS transistor. The metal-nitride-oxide-silicon (MNOS) transistor holds a particular place in the family of MOS transistors. In this transistor, the dielectric has a "sandwiched" structure consisting of the layers of silicon nitride Si_3N_4 and silicon dioxide (Fig. 7.33a). The oxide layer 2 to 5 nm thick is formed by thermal

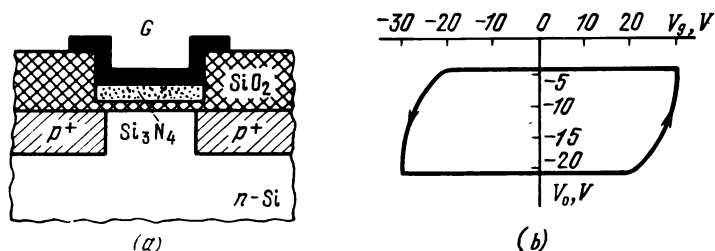


Fig. 7.33. Induced p -channel MNOS transistor
(a) structure; (b) threshold voltage versus gate voltage

oxidation, and the nitride layer by reactive cathode sputtering. The Si_3N_4 layer has a thickness of 0.05 to 0.1 μm . This is sufficient for the breakdown voltage to be in excess of 50 to 70 V.

The main feature of a MNOS transistor is that its threshold voltage can be varied by applying to the gate short pulses (100 μs) of different polarity and of large amplitude (30 to 50 V). Thus the application of +30 V can set up a threshold voltage V_0 of -4 V (Fig. 7.33b). This value remains stable in the further operation of the transistor when used in the small-signal mode ($V_g \leq \pm 10$ V); the transistor acts then as an induced p -channel MOST. If we now apply a pulse of -30 V, the threshold voltage V_0 will become equal to -20 V. In this condition, the same signals $V_g = \pm 10$ V will fail to render the transistor conducting. As is apparent, the *hysteretic* $V_0 = V_g$ relation permits the MNOS transistor to be switched from the on to the off state and vice versa by applying sufficiently large control pulses. This feature is put to use in integrated memories.

The MNOS transistor operates on the principle of charge storage at the boundary between the nitride and the oxide layer. This storage results from the difference between conduction currents in each of the layers. The charge buildup obeys the elementary equation

$$dQ/dt = I_{\text{SiO}_2} - I_{\text{Si}_3\text{N}_4}$$

where both currents depend on the gate voltage and vary during charge storage. At a high **negative** voltage V_g , a positive charge builds up at the boundary. This process is equivalent to the introduction of donors into the dielectric and entails an increase in the negative threshold voltage. With a large **positive** voltage V_g applied to the gate, a negative charge develops at the boundary, which decreases the negative threshold voltage.

At low voltages V_g , the currents in dielectric layers decrease by factors between 10^{10} and 10^{15} (!), so that *the induced charge and the corresponding threshold voltage can be retained for thousands of hours.*

7.9. Semiconductor Resistors

The early semiconductor ICs used only *diffused resistors* (DR), the main constituent of which was one of the **diffused** layers disposed in an isolated island. In common use now are also ion-implanted resistors.

7.9.1. Diffused resistors. This type of resistor is most commonly formed by a base layer stripe with two ohmic contacts (Fig. 7.34). For this stripe configuration, the resistance of a DR according to Eq. (7.4) can be written in the form

$$R = R_s (a/b) \quad (7.5a)$$

where R_s is the sheet resistance (see p. 223-4); the dimensions a and b are shown in Fig. 7.34.

Both the length and width of a striped DR are limited. The length a cannot be larger than the crystal size, and thus lies in the range from 1 to 5 μm . The factors that set the lower limit to the width b are the capabilities of photolithography, lateral diffusion, and permissible spread (10 to 20%). The minimum width is generally 10 to 15 μm .

Substituting $R_s = 200$ ohms/square and $a/b = 100$ in Eq. (7.5a) gives a typical maximum value of $R_{\max} = 20$ k Ω . This value can be increased twofold or threefold using a *meandered configuration* (Fig. 7.35). The resistance of a meandered resistor assumes a more

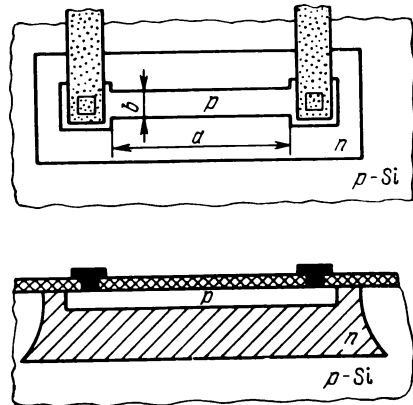


Fig. 7.34. Stripe-type configuration of a diffused resistor

general form

$$R = R_s \left(\frac{\sum_i a_i}{b} + n + 1.3 \right) \quad (7.5b)$$

Here n is the number of bends (in Fig. 7.35, $n = 2$); and the summand 1.3 is a factor that takes into account the resistor inhomogeneity in the region of ohmic contacts.

The number of bends, n , is of course limited by the area assigned to the diffused resistor and does not commonly exceed 3. Otherwise the resistor area can reach 15 to 20% of the entire chip area. The maximum resistance at $n = 3$ does not exceed 50 to 60 k Ω .

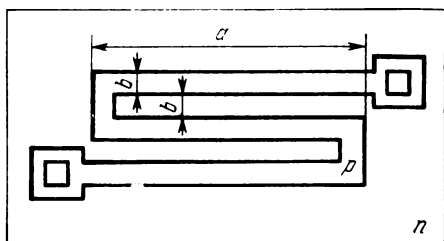


Fig. 7.35. Meandered-type configuration of a diffused resistor

The TCR of a base-diffusion resistor ranges from 0.15 to 0.30 % $^{\circ}\text{C}^{-1}$ depending on the value of R_s . The spread in resistances about the rating is between ± 15 and ± 20 %; the values of resistors formed on the same chip change in one direction only, for which reason the ratio between resi-

stances is kept to a closer tolerance, ± 3 % and below, and the TC for the ratio between resistances does not exceed ± 0.01 % $^{\circ}\text{C}^{-1}$. This feature of a diffused resistor plays an important role and finds wide use in the development of ICs.

Where the desired resistance ratings exceed 50 to 60 k Ω , it is possible to use the so-called *pinch resistor* of the structure as shown in Fig. 7.36. In comparison with the simplest diffused resistor of Fig. 7.34, the pinch resistor has a smaller cross section and a larger sheet resistance (owing to the use of the weakly doped, bottom portion of the p layer). As a result, the R_s value of a pinch resistor commonly reaches 2 to 5 kilohms per square and over, depending on the thickness. At such a value of R_s , the maximum resistance can be as high as 200 to 300 k Ω even for the simplest striped configuration.

Shortcomings of a pinch resistor are a large tolerance for rated values (up to 50 %) because of a thin p layer, a higher TCR (0.3 to 0.5 % $^{\circ}\text{C}^{-1}$) due to a low level of doping at the bottom of the p layer, and nonlinearity of the I - V characteristic at voltages above 1 to 1.5 V. The last shortcoming stems from the analogy between the structures of a pinch resistor and a field-effect transistor (see Fig. 7.26b). The I - V curve of a pinch resistor coincides with that of a JFET (see Fig. 5.14a) if the JFET gate voltage is set to zero (because the pinch resistor has its n^+ and p layers connected with each other by

metallization). The breakdown voltage of a pinch resistor depends on that of the emitter junction, and generally ranges between 5 and 7 V.

If the desired resistance ratings must be 100 Ω and below, the use of the base layer in a diffused resistor is unsuitable because the width of the resistor must then be smaller than its length, see Eq. (7.5a), which is difficult to implement constructionally. To produce a low-value diffused resistor, use is made of a low-resistance emitter layer. At R_s equal to 5 to 15 ohms/square, which values are typical of this layer (see Table 7.1), it becomes possible to obtain

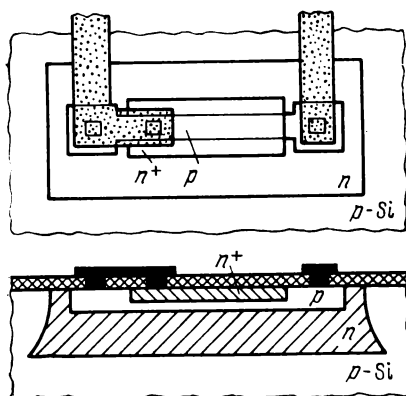


Fig. 7.36. Pinch resistor

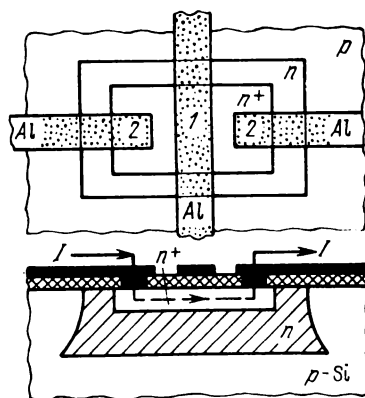


Fig. 7.37. Underpass

minimum resistances of 3 to 5 Ω with a TCR of 0.01 to 0.02% $^{\circ}\text{C}^{-1}$.

Using the emitter n^+ layer permits solving one more problem encountered in designing an IC, namely, implementing a *tunnel crossing*, or *underpass* (Fig. 7.37). The question is how to isolate two mutually perpendicular metallization stripes of which the first 1 lies on the protective oxide and the second, 2, partially passes under the first; this "crossunder" portion represents a low-resistance n^+ region. An example of the tunnel crossing can be a portion of the collector n^+ layer shown in Fig. 7.13b. The underpass is not a universal solution because the n^+ region yet has a noticeable resistance, typically 3 to 5 Ω . Therefore, *underpasses are unsuitable for, say, supply circuits* carrying rather high currents.

7.9.2. Ion-doped resistors. Last years have seen ever wider use of *ion-doped resistors* produced by local ion implantation (see Subsec. 6.5.3) rather than by diffusion as is the case for diffused resistors.

The structure of an ion-doped resistor is the same as that of a diffused resistor (see Fig. 7.38), but the implantation depth of the p layer is much lower than for the base layer and is equal to merely

0.2 or 0.3 μm . Besides, ion implantation can provide as small an impurity concentration in the layer as desired. Both factors favor the formation of a layer of high sheet resistance, up to 10-20 kilohms/square; the resistance ratings can range into hundreds of kilohms.

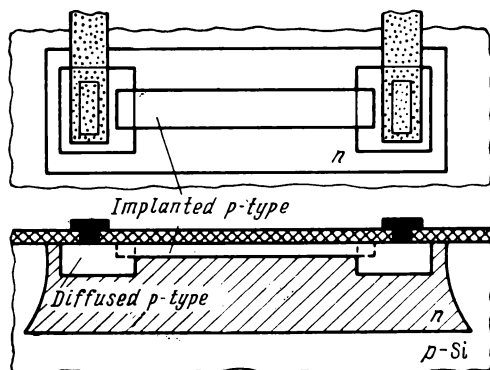


Fig. 7.38. Ion-doped resistor

The TCR is lower than for diffused resistors and ranges 3 to 5 % $^{\circ}\text{C}^{-1}$; the spread in resistances does not exceed ± 5 to ± 10 %.

Since the thickness of an implanted layer is small, it is difficult to accomplish ohmic contacts to this thin layer. The approach therefore is to diffuse narrow p layers at the resistive layer edges during the base diffusion and then to apply an ohmic contact to these diffused layers by the common method.

7.9.3. Equivalent circuits. A characteristic feature of any integrated resistor is that it shows a *parasitic capacitance* to the substrate

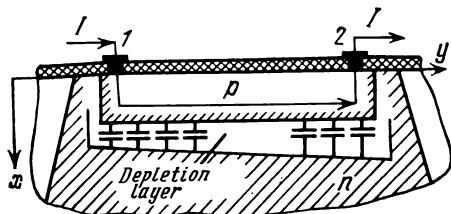


Fig. 7.39. Physical model of an integrated resistor in the form of a distribute RC circuit

or isolating well. In the simplest DR (see Fig. 7.34), such a *parasitic capacitance* is a barrier capacitance on the junction formed between the working p layer and n epi-layer of the island¹.

Strictly speaking, the combination of the resistance and parasitic capacitance represents a distributed RC circuit (Fig. 7.39). In ap-

¹ The n layer is at the most positive voltage of the power source, therefore all the portions of the pn junction are at the reverse voltage.

proximate calculations, however, it is more convenient to use equivalent circuits with **lumped** constants, such as a Π equivalent circuit (Fig. 7.40a) or T equivalent circuit (Fig. 7.40b). In these circuit models, R is the value of the resistor, and C is the mean junction capacitance.

The necessity for averaging the capacitance arises from the following fact. As the current flows across the resistor, the potential

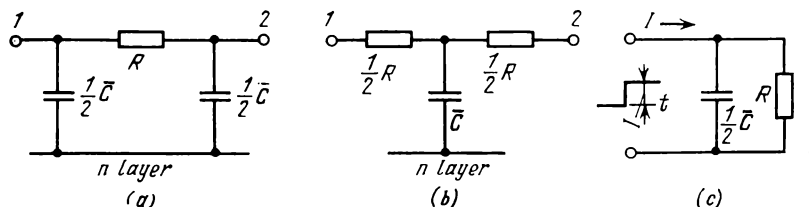


Fig. 7.40. Equivalent circuits of an integrated resistor
(a) Π circuit; (b) T circuit; (c) Π circuit at $V_2 = \text{const}$

on the p layer becomes different at different points. Since the potential on the n layer is constant, the potential difference across the pn junction will vary along the y -axis, and hence the barrier capacitance will vary too.

In a typical case where one of the resistor terminals, say, terminal 2 is at the constant potential, while the other terminal 1 is used to apply a step of current, the Π circuit reduces to the simplest RC circuit shown in Fig. 7.40c. The transient response essentially occurs as a smooth change of the voltage of the resistor with a step-like change of the current. The time constant that determines the transient has the form

$$\tau = 1/2 R \bar{C} \quad (7.6a)$$

and the corresponding cutoff frequency

$$f_{cut} = 1/2 (\pi \tau) = 1/(\pi R \bar{C}) \quad (7.6b)$$

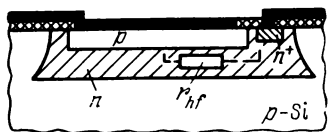
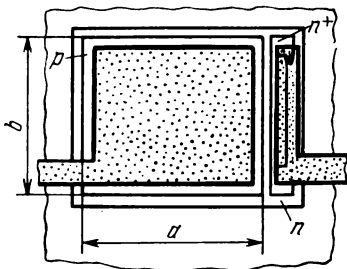
If $R = 10 \text{ k}\Omega$ and $\bar{C} = 1.3 \text{ pF}$, then $\tau = 6.5 \text{ ns}$ and $f_{cut} \approx 25 \text{ MHz}$. This means that in the given example the resistor performs its function as a purely resistive element only at frequencies up to 10 to 15 MHz. At higher frequencies, the reactance comes into play, and so the operation of the circuit using this resistor can change substantially.

The described equivalent circuits are also valid for other versions of resistors: where the emitter or collector layers are used as a working region, and also where use is made of oxide isolation of elements. The quantitative results, however, become different. With the use of oxide isolation, for example, the time constant can be a few times smaller.

7.10. Semiconductor Capacitors

In bipolar semiconductor ICs, it is reverse-biased pn junctions that act as monolithic capacitors. In these capacitors, at least one layer is of the diffused type, for which reason they received the name of *diffused capacitors*.

7.10.1. Diffused capacitor. The typical structure of a diffused capacitor where the collector-base junction serves as a capacitive element is illustrated in Fig. 7.41. The capacitance of such a capacitor has the general form



$$C = C_{01}(ab) + C_{02}2(a + b)d \quad (7.7a)$$

where C_{01} and C_{02} are the per-unit area capacitances of the bottom and side portions of the pn junction respectively. The relation between the summands on the right of Eq. (7.7a) depends on the ratio a/b , that is, on the configuration of a diffused capacitor. *The optimal geometry is the square* ($a = b$): here the "side" component of capacitance is tens of times smaller than the bottom component. Neglecting the side component, that is, the addend in (7.7a), and assuming $a = b$, we get

$$C = C_{01}(ab) = C_{01}a^2 \quad (7.7b)$$

Fig. 7.41. Diffused capacitor

For example, if $C_{01} = 150 \text{ pF/mm}^2$ and $C = 100 \text{ pF}$, then $a \approx 0.8 \text{ mm}$. As seen,

the size of the capacitor turns out to be **comparable** to the chip size.

In order that the total area of all capacitors entering into the IC might not be in excess of 20 to 25% of the chip area, we must limit the total capacitance of capacitors to the quantity

$$C_{\max} = (0.2 \text{ to } 0.25) C_{01}S_{ch}$$

where S_{ch} is the useful chip area. If $S_{ch} = 2 \text{ to } 5 \text{ mm}^2$ and $C_{01} = 150 \text{ pF/mm}^2$, then $C_{\max} = 50 \text{ to } 200 \text{ pF}$.

Using an emitter rather than a collector pn junction can result in a 5-fold increase in maximum capacitance because of a larger per-unit area capacitance of the emitter junction.

The basic parameters of a diffused capacitor, including the spread in nominal values, δ , due to manufacturing factors, the TCC¹,

¹ The dependence of the capacitance of a diffused capacitor on temperature stems from the function $\Delta\varphi_0(T)$ [see Eq. (3.4)], which enters into Eqs. (3.34).

the breakdown voltage V_{br} , and Q factor appear in Table 7.4 for both versions of a diffused capacitor using the collector or the emitter junction. As apparent from the Table, the emitter junction capacitor

Table 7.4

Typical Parameters of Integrated Capacitors

Capacitor type	C_0 , pF/mm ²	C_{max} , pF	δ , %	TCC , % °C ⁻¹	V_{br} , V	Q (1 MHz)
Collector junction	150	300	± 20	-0.1	50	50-100
Emitter junction	1 000	1 200	± 20	-0.1	7	1-20
MOS structure	300	500	± 25	0.02	20	200

offers higher maximum capacitances, which is its major advantage. As regards the breakdown voltage and Q , this version is inferior to the collector junction counterpart.

The required condition for the normal function of a diffused capacitor is the **reverse biasing** of its pn junction. So *the voltage at the DC must have a strictly definite polarity*.

The capacitance of diffused capacitors according to Eq. (3.34) depends on voltage. This means that a diffused capacitor is generally a *nonlinear capacitor* with a C - V characteristic as shown in Fig. 3.17. Nonlinear capacitors perform useful functions in special units of radio engineering equipment, such as parametric amplifiers and frequency multipliers. But in more extensive use are *linear* capacitors with a **constant** capacitance, which are able to pass ac components of the signal without distortion and block (hold) dc components. The diffused capacitor successfully handles this function **under the constant bias** E that exceeds the amplitude of an ac signal.

Let, for example, the total reverse voltage impressed on the diffused capacitor be of the form

$$V = E + V_m \sin \omega t$$

where $E = \text{constant}$ and $V_m \ll E$. In this case the voltage V varies in the range from $E + V_m$ to $E - V_m$, that is, changes insignificantly. The capacitance of the diffused capacitor thus practically remains constant and equal to $C(E)$. The ac component of current will be determined by the capacitive susceptance as in the "common" capacitor:

$$I_m = Y_C V_m$$

where $Y_C = \omega \cdot C(E)$.

An important feature of the diffused capacitor is the possibility of varying the value of capacitance by changing the bias E . Conse-

quently, the diffused capacitor can serve not only as a traditional capacitor but also as a *capacitor with an electrically controlled capacitance*, or as a *variable capacitor*. Such capacitors are useful for, say, adjustment of tuned circuits. The electrical adjustment of capacitance is of course more preferable than the mechanical one. But the range of electrical adjustment is rather narrow: varying the bias E from 1 to 10 V, we can change the capacitance of a diffused capacitor by no more than a factor of 2 to 2.5 [see Eqs. (3.34)].

7.10.2. Quality factor. An important parameter of any capacitor, the diffused capacitor included, is a *high-frequency* Q_{hf} . The Q_{hf} factor characterizes the power loss with the passage of capacitive current and is defined as a ratio

$$Q = \frac{X_C}{r_{hf}} = \frac{1}{\omega C r_{hf}} \quad (7.8a)$$

where r_{hf} is the loss resistance at high frequencies (Fig. 7.42a). The lower the active power as against the reactive power, the higher

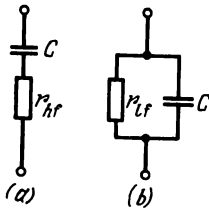


Fig. 7.42. Physical models of diffused capacitors for hf (a) and for lf (b)

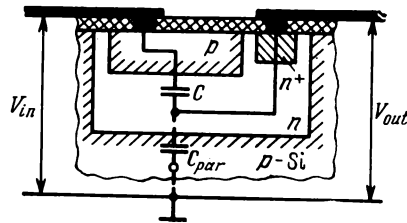


Fig. 7.43. Role of parasitic capacitance in conveying ac voltage across a diffuse capacitor

the Q factor. For example, if $C = 100$ pF, $r_{hf} = 20 \Omega$ and $f = 1$ MHz, then $Q_{hf} \approx 75$. In an ideal capacitor, $r_{hf} = 0$ and $Q_{hf} = \infty$.

The main source of loss in a diffused capacitor is the lateral resistance of lower layers that enter into pn junction structure. For a collector junction, this is the series layer resistance (see Fig. 7.41); for an emitter junction, the source of loss is the base layer resistance. With the n^+ buried layer being present, the r_{hf} for the collector junction is much lower than for the emitter junction. Therefore, the Q for a diffused capacitor is lower when using the emitter junction rather than the collector junction (see Table 7.4).

From Eq. (7.8a) it follows that the Q grows with decreasing frequency. At rather low frequencies, however, another type of loss makes itself felt to a considerable degree, which is quite negligible

at high frequency. It is the **leakage resistance** r_{lf} that stems from the reverse current through a *pn* junction. This resistance shunts the capacitance of a diffused capacitor (Fig. 7.42b). The *quality factor at low frequencies* is therefore the ratio

$$Q_{hf} = \frac{Y_C}{1/r_{lf}} = \omega C r_{lf} \quad (7.8b)$$

where r_{lf} is the leakage resistance at low frequencies.

The resistance r_{lf} is inversely proportional to the junction area since it is a function of the junction reverse current. So the product $C r_{lf}$ and hence low-frequency quality factor are independent of the area. The Q_{lf} at 500 Hz is typically 50 to 100.

Over the frequency range in which both the Q_{hf} and Q_{lf} exceed 100 to 200, the diffused capacitor represents an almost ideal capacitor, so the equivalent circuits (see Fig. 7.42) can disregard the loss resistance. The above examples show that the diffused capacitor is nearly ideal in the frequency range from 500 Hz to 500 kHz.

7.10.3. Equivalent circuit. A specific feature of the diffused capacitor as an integrated element is that it exhibits a **parasitic capacitance**. In a collector junction capacitor, this is a barrier capacitance between the collector layer and substrate, $C_{par} = C_{c\ sub}$ (see Fig. 7.14). *The parasitic capacitance prevents the diffused capacitor from transferring the applied input voltage fully to the load.*

Indeed, from the equivalent circuit of Fig. 7.43 it is apparent that the capacitance of a diffused capacitor together with the parasitic capacitance forms a capacitive voltage divider, so only a **fraction** of input voltage V_{in} appears at the output:

$$V_{out} = V_{in} \frac{X_{par}}{X_{par} + X_C}$$

where X_{par} and X_C are the reactances of the parasitic and working capacitances, C_{par} and C , respectively. Substituting $X_{par} = 1/\omega C_{par}$ and $X_C = 1/\omega C$, we may write the voltage transfer ratio as

$$V_{out}/V_{in} = C/(C + C_{par}) \quad (7.9)$$

The voltage ratio will be close to unity if the inequality $C_{par} \ll C$ is valid. But the areas of both capacitors (working and parasitic) are almost equal (see Fig. 7.43). Moreover, the area of the parasitic capacitor $C_{c\ sub}$ is even a little larger.

Therefore, the capacitances C_{par} and C differ only on account of the difference between the **per-unit area** capacitances of collector and emitter junctions and also as a result of the difference between the **voltages** applied to these junctions. Calculations show that in real IC structures it is hardly possible to decrease C_{par} below 0.15 to 0.2 C . Hence the voltage transfer ratio lies between 0.8 and 0.9.

Essentially the same conclusions and equivalent circuit hold for the diffused capacitor using an emitter junction.

7.10.4. MOS capacitor. This integrated capacitor is principally different from the diffused capacitor. Fig. 7.44 shows its typical structure where a thin oxide layer 0.08 to 0.12 μm thick is grown

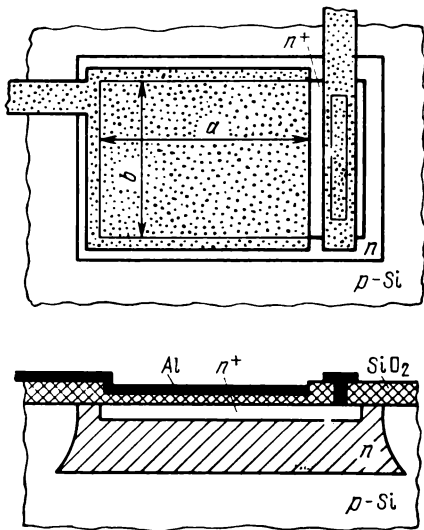


Fig. 7.44. MOS capacitor having SiO_2 as dielectric

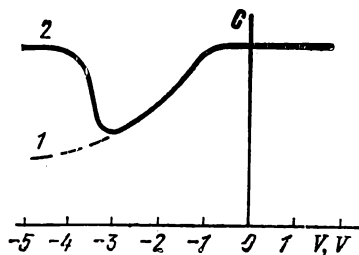


Fig. 7.45. C - V characteristic of a MOS capacitor

1—inversion layer absent; 2—inversion layer present

surface region becomes **enriched** with electrons at the zero and positive voltages on the metal plate, and hence the depletion layer is absent. The total capacitance is thus determined by the dielectric and has a maximum value.

by an additional manufacturing process above the emitter n^+ layer. In the subsequent operation, while depositing the metalization pattern, this thin layer is coated with aluminum to produce an upper plate of the capacitor. The emitter n^+ layer acts as a lower plate.

The per-unit area capacitance of a MOS capacitor is expressed by Eq. (7.4) and is typically equal to about 350 pF/mm^2 . The basic parameters of a MOS capacitor are given in Table 7.4.

An important advantage of the MOS capacitor over the diffused capacitor is that the former can operate with **any** polarity of voltage. In this respect, this capacitor is analogous to a traditional capacitor. But the MOS capacitor, like the diffused capacitor, is nonlinear; an example of the C - V characteristic appears in Fig. 7.45.

The C - V relation results from the fact that the capacitance of a MOS capacitor generally represents two capacitances connected **in series**: the dielectric capacitance mentioned earlier and the depletion layer capacitance that can form in the surface region of the semiconductor. In a capacitor shown in Fig. 7.44, the

At negative voltages, a depletion layer gradually builds up, the depth of which grows with voltage. The depletion layer capacitance decreases accordingly, leading to a decrease in the total capacitance of the MOS capacitor (curve 1). At a rather high negative voltage, an inversion p layer (conducting channel) appears near the surface. Thus the depletion layer capacitance becomes "disconnected" from the dielectric capacitance, and the total capacitance of the MOS capacitor approaches the initial value (curve 2).

For the effect of the depletion layer to be insignificant, it is necessary that the capacitance of this layer should be large as against the capacitance of the dielectric. This requirement can be fulfilled with a large impurity concentration in the semiconductor. It is for this reason that the n^+ layer is used as a semiconducting "plate" in the MOS capacitor. A low resistance of this layer also ensures a high Q factor of the capacitor.

The parasitic capacitance C_{par} in the equivalent circuit of Fig. 7.43 should be understood as the capacitance between the n island and p substrate. The voltage transfer ratio of Eq. (7.9) is not lower than 0.9 to 0.95.

A distinguishing feature of the MOS capacitor is that its capacitance is frequency dependent. This dependence arises from the effect of fast surface states at the semiconductor-dielectric boundary. The recharge of these states is an inertial process, which proceeds at a time constant of about $0.1 \mu s$ (see p. 36). That is why an increase in frequency makes the capacitance of a MOS capacitor drop off and reach a steady value only at frequencies above a few megahertz.

Let us note in conclusion that in distinction to bipolar ICs, the fabrication of MOS capacitors in MOS ICs does not involve extra operations: the thin oxide for capacitors is formed at the same stage as that used for growing a thin oxide under the gate, and the low-resistance semiconductor layer at the stage of the source and drain doping.

7.11. Film Integrated Elements

As known, the commercial fabrication of active film elements such as diodes and transistors has yet to be developed. Therefore we consider only passive film elements such as resistors, capacitors, and inductors. These elements can be made by both the thin film and the thick film technology. The configurations of thin film and thick film elements are identical, but they can differ substantially in geometrical dimensions (at given electrical parameters) because of the use of entirely different materials.

Film elements need not be mutually isolated because they are deposited on the dielectric substrate. Because the substrate is comparatively thick, not less than $500 \mu m$, and the distances between

elements are rather large, parasitic capacitances are insignificant and not included into equivalent circuits.

7.11.1. Resistors. The structure and configuration of a film resistor appears in Fig. 7.46. In the general case, the configuration of a film resistor is the same as for the diffused resistor shown in Fig. 7.35. As seen, the configuration is of the meandered type, though

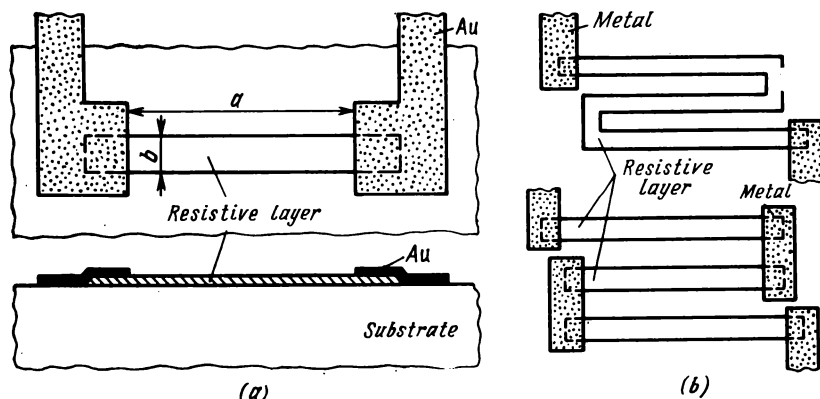


Fig. 7.46. Film resistors of the striped configuration (a) and meandered configuration (b)

stripe-type resistors are also available. So, Eq. (7.5) is applicable here for the calculation of resistances. The sheet resistance depends on layer thickness and the material used. The typical values of R_s are listed in Table 7.5, which also includes the typical values of

Table 7.5

Typical Parameters of Film Resistors

Resistor type	$R_s, \Omega/\square$	R_{\max}, Ω	R_{\min}, Ω	$\delta, \%$		TC, $10^{-4}/^{\circ}\text{C}$		$\Delta R(t), \%$ (1 000 h, 70°C)
				not trim- med	trimmed	R	R_1/R_2	
Thin film	10-300	10^6	10	± 5	± 0.05	0.25	± 0.05	0.005
Thick film	$5.0 \cdot 10^6$	5×10^8	0.5	± 15	± 0.2	2	± 0.1	0.05

other parameters such as the maximum and minimum resistance ratings, resistance tolerances δ , TCR, and resistance drift (for 1 000 h at 70°C).

The table gives the tolerances, δ , for two cases; the first refers to resistors not subjected to special adjustment, or trimming, and

the second to resistors subjected to trimming. The methods of trimming will be discussed later. The values of TCR in the table are also given for two cases: for the resistance of an individual resistor, R , and for the resistance ratio of two resistors, R_1/R_2 .

The data of Table 7.5 allows us to make the following general conclusions.

1. Film resistors span by far a wider resistance range than monolithic counterparts such as diffused and ion-implanted resistors.
2. Thin film technology produces resistors of higher precision and stability.

3. Trimming substantially decreases the spread in resistance. So the possibility of trimming the resistor values is an important advantage of film resistors.

4. The resistance ratio features a smaller spread and lower TCR than an individual resistance (this is also the case for monolithic resistive elements).

A few methods are available for trimming the resistors. The simplest and historically first method consists in partial mechanical scraping of the resistive layer prior to depositing a protective coat on the surface of the IC substrate. A more advanced approach comes to a partial removal of the layer with an electric spark, electron or laser beam. These methods eventually raise the resistance. The most perfect and flexible method consists in passing a rather high current through a resistor. This method initiates two simultaneous processes: oxidation of the resistive layer surface and ordering of its fine-grained structure. The first process tends to raise the resistance and the second to decrease it. Selecting the proper value of current and the atmosphere most suitable for trimming can ensure a change in resistance in either direction by $\pm 30\%$ to an accuracy of fractions of a percent with respect to the desired rated value.

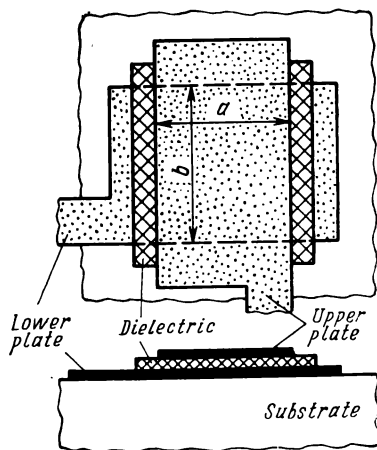


Fig. 7.47. Film capacitor

7.11.2. Capacitors. Fig. 7.47 illustrates the structure and configuration of a typical film capacitor whose per-unit area capacitance is found from Eq. (7.4). The thickness d of a dielectric film strongly depends on the technology employed: for thin films, $d = 0.1$ or $0.2 \mu\text{m}$, and for thick films, $d = 10$ to $20 \mu\text{m}$. Other things being the same, the per-unit area capacitance of thick film capacitors is

smaller than of thin film ones. However, the difference in dielectric thickness can be offset owing to the difference in the permittivities of materials.

In thin film capacitors, the per-unit area capacitance is not proportional to the permittivity of the material used because the breakdown strength also plays its role. The material with a high value of ϵ can have a low breakdown strength. At a given breakdown voltage, therefore, one should increase the dielectric layer thickness; thus the gain in per-unit area capacitance proves smaller than would be expected.

When making a choice on the dielectric for a **high-frequency** capacitor (both a thin film and a thick film type), we should also consider the loss of energy in the dielectric¹.

As for the ohmic loss in the plates of film capacitors, this is much lower than for monolithic capacitors, because both plates are metallic layers of high conductance.

Table 7.6

Typical Parameters for Film Capacitors

Capacitor type		C_0 , pF/mm ²	C_{\max} , pF ($S = 25$ mm ²)	δ , %	TCC, %°C ⁻¹	Q (10 MHz)
Thin film	SiO	60	1 500	± 15	0.2	200
	Al ₂ O ₃	1 500	4×10^4	± 15	0.03	30
	Ta ₂ O ₅	4 000	10^5	± 15	0.02	30
Thick film		—	10^4	± 20	to ± 0.05 ± 0.15	—
MOS		350	200	± 20	0.02	10

Table 7.6 gives the typical parameters of film capacitors and also shows for comparison the parameters of MOS capacitors which resemble the former in structure. What can be inferred from the table is as follows.

¹ This power loss is described by $\tan \delta$, which is the tangent of the angle δ formed between the phasor of total current and the phasor of reactive current component through a capacitor at a given frequency. If the loss is insignificant, $\tan \delta \approx \delta \ll 1$.

1. The per-unit area capacitance of film capacitors, with the dielectric being properly chosen, can be ten times as high as that of MOS capacitors, let alone diffused capacitors.

2. Maximum capacitances of film capacitors can be a few orders of magnitude higher than those of monolithic capacitors, largely because of an increased area; the fact is that the area of hybrid IC substrates greatly exceeds the area of semiconductor IC chips.

3. Thick film capacitors are almost comparable to thin film capacitors in most of the parameters, the temperature coefficient being probably the exception.

4. For high-frequency thin film capacitors, the optimal dielectric is silicon monoxide. Germanium monoxide parameters resemble those of silicon monoxide.

It should be pointed out that miniature discrete capacitors are now available, including the ones exhibiting rather a high capacitance, typically up to a few microfarads. That is why there is a tendency today to replace film capacitors by chip capacitors.

7.11.3. Inductors. As noted earlier, the possibility of the manufacture of inductive components by microelectronic techniques is one of the merits of film technology. These components have the shape of a flat square or round spiral, the former configuration being most popular (Fig. 7.48). The material used is largely gold since it has low resistivity. The width of a metal stripe is 30 to 50 μm , and the gap between turns is 50 to 100 μm . With such a geometry, the per-unit area inductance can range from 10 to 20 nH/mm², so a surface area of 25 mm² can give an inductance of 250 to 500 nH.

The Q_{hf} of a film spiral inductor is defined by relation

$$Q_{hf} = \omega L / r_{hf} \quad (7.10)$$

where r_{hf} is the loss resistance at high frequencies. Thus at 100 MHz the quality factor can be $Q_{hf} = 50$. Unlike the Q of a capacitor [see Eq. (7.8a)], the Q of an inductor rises with frequency. For this reason film inductors containing 3 to 5 turns can operate to advantage in the SHF range at 3 to 5 GHz.

The advent of miniature fine-wire wound coils has limited the use of film inductors, particularly at frequencies below 50-100 MHz; because it is advantageous to use discretes in preference to film inductors, as is the case with capacitors.

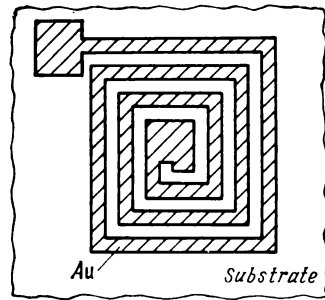


Fig. 7.48. Film square-spiral inductor

8.1. General

The advancement of microelectronics inevitably involved a kind of “natural selection” of most suitable circuits from a large number of electronic circuits employed earlier in discrete transistor electronics. A large amount of circuits, which were typical and most popular in discrete version, proved far from being optimal or, sometimes, impractical in integrated circuit form. On the contrary, the circuits which found rare and rather narrow special-purpose applications in discrete electronics have held the lead in microelectronics. Besides, microelectronics, as a qualitatively new stage in electronic engineering, has given birth to a number of new, unknown earlier, circuit designs whose realization in discrete circuit form could be impossible.

At the present time, there is no sense in studying electronic circuit engineering within such a scope, in such detailed form, and even in such a sequence as it was justifiable only 10 to 15 years back. The former kind of presentation would largely be encyclopediac in character and, in some cases, should have to entail negative estimates and recommendations as regards the practical use of one circuit or another in integrated version. The circuits discussed in Chapter 8 and 9 represent that “gold stock” which has passed from discrete transistor engineering to microelectronics and formed the basis for its further progress.

8.2. Classification of Electronic Circuits

At present, it is common to divide electronic circuits into two classes, *digital* and *analog*.

Digital circuits mainly use the simplest *transistor switches*, which are analogs of metallic contacts. Switches feature two steady states, open (off) and closed (on). The simplest switches underlie the structure of more complex circuits such as logic, bistable, trigger, and other types.

Analog circuits rely on the simplest amplifying *units* or *stages*. Amplifying stages make up complex (multistage) amplifiers, current and voltage regulators, frequency converters (modulators and discriminators), sinusoidal oscillators, and a number of other circuits. Such circuits sometimes go under the name of linear or quasilinear

circuits, though this name is more appropriate to amplifiers and stabilizers, while for other circuits it is rather conditional.

It is convenient to explain the specifics of digital and analog circuits with *transfer characteristics* describing the variation of an output variable with an input variable. For example, we assume that these variables are voltages.

Figure 8.1 illustrates typical transfer characteristics. Curve 1 shows the characteristic of *inverting* circuits in which low input voltages correspond to high output voltages, and curve 2 shows the characteristic of *noninverting* circuits in which low input voltages correspond to low output voltages. Inverting circuits find more extensive uses. The transfer characteristics are also typical of the simplest switches and the simplest amplifying stages. But the use of these characteristics in both classes of circuits is principally different.

In a transistor switch, its two steady states, the off and the on state, correspond to points A and B shown in Fig. 8.1. At point A, the switch is off, so the voltage drop across it is high. At point B, the switch is on and the voltage drop across it is close to zero. Input and output signals (voltages) in the switch assume only two values: either V_{inA} and V_{outA} or V_{inB} and V_{outB} . The shape of the curve between the points A and B is of no significance; if it changes as shown by a dash line, the output signals remain practically invariable. From this it follows that simple switches and hence digital circuits are little sensitive to the spread in parameters, to thermal and time drift, and also to external electromagnetic interference (stray pickup) and intrinsic noise. Fig. 8.1 illustrates the last conclusion by showing that small voltage variations ΔV_B about point B (such as inherent or external noise) do not practically alter the value of output signal and thus have no effect on the switch operation.

The amplifying stage operates in the region between points a and b. The input and output signals may take on any values within this region, and the functional relation between the input and output has the form: $V_{out} = f(V_{in})$. Obviously, any "deformation" of the characteristics in the a-b region, whatever the reason of this deformation can be, will affect directly the above function and hence the operation of the circuit. For example, at the same input signal V_{inC} , the output signal may take different values, either $V_{outC'}$ or $V_{outC''}$. What can be inferred from the above is that an amplif-

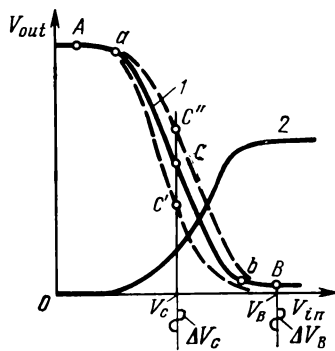


Fig. 8.1. Transfer characteristics 1 and 2 of inverting and noninverting circuits respectively

ying stage and thus analog circuits are responsive to the spread in parameters, to the thermal and time drift, and also to noise and stray pickup. As seen from Fig. 8.1, small voltage variations ΔV_C about point C cause a noticeable change in the output signal in accordance with the function $V_{out} = f(V_{in})$.

Each type of electronic circuit, whether digital or analog, naturally has a more detailed classification, first of all, by the functions they have to perform. We shall start the study of microelectronic circuitry principles by considering transistor switches and digital circuits, because of their relative simplicity, though analog circuits were the first to appear in electronics; it was these circuits that formed the foundation for radio engineering—the first field of industry that began to use electronic devices. Analog circuits will be discussed in the next chapter.

8.3. Static Operation of a Simple Bipolar Switch

In the static operation of a switch, that is, at two stable points A and B shown in Fig. 8.1, it is metallic contacts that feature ideal parameters. In these mechanically operated contacts the residual

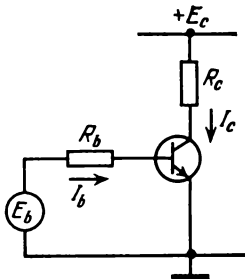


Fig. 8.2. Simple transistor switch

current in the off condition depends on the quality of insulation, and does not commonly exceed 10^{-12} A. In the on condition, the residual voltage at the contact comes to fractions of a microvolt at currents of about 1 mA. By these parameters, mechanical switches still remain beyond competition.

But in the dynamic operation, that is, in switching from one operating point to the other, mechanical switches are much inferior to electronic counterparts in maximum switching frequency, contact reliability, and service life. These characteristics have predetermined the choice of electronic switches in preference to mechanical ones for use in digital circuitry.

8.3.1. Operating points. Fig. 8.2 shows the circuit of the simplest transistor switch. The transistor is connected in a common emitter configuration. The collector circuit that includes a power source E_c and load resistor R_c is a controlled (switched) circuit. The controlling (base) circuit includes a control voltage source E_b and a series resistance R_b .

If the voltage E_b is negative, the emitter junction is reverse biased, the transistor is turned off, and so the residual current in the load circuit is very small. The voltage V_{ce} across the switch is thus close to E_c .

With the voltage E_b being positive and having a sufficiently high value, the transistor is in the on condition, the current I_c flows in the load circuit, and the residual voltage across the switch can be close to zero.

From the foregoing description we can conclude that the switch is an inverting circuit because an increase in the input voltage

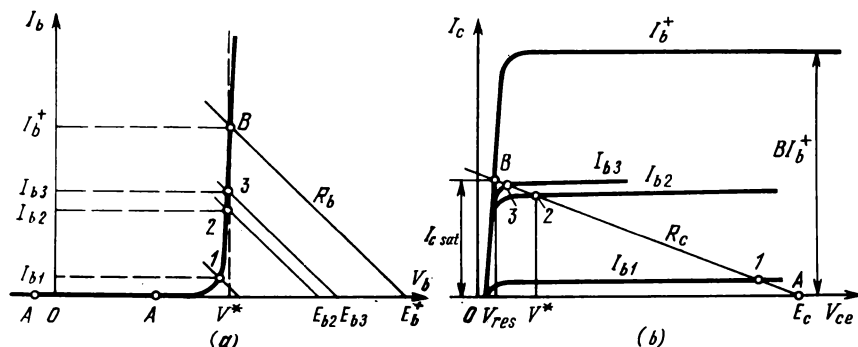


Fig. 8.3. Operating points on the input (a) and output (b) static characteristics of a switch

E_b from negative to positive values entails a decrease in the output voltage V_{ce} from E_c to a small residual voltage.

The residual current and residual voltage are the main static parameters of the switch. Consider them in more detail.

A transistor switch in the off state must, strictly speaking, obey the condition $E_b < 0$. However, the silicon *pn* junction remains practically off also at a small positive bias voltage $E_b < 0.6$ V (see p. 89). In the off state, the currents at the three terminals of the transistor do not generally exceed fractions of a microampere. We thus can neglect the voltage drop across resistances R_b and R_c and set $V_b = E_b$ and $V_{ce} = E_c$. The point A on Fig. 8.3 represents the off condition of the switch.

When the voltage E_b reaches V^* , the transistor becomes conductive. The base current I_{b1} starts to flow and so does the collector current I_{c1} in proportion to I_{b1} ; the collector potential correspondingly decreases, as shown at points 1 in Fig. 8.3. As E_b grows further, the base potential V_b remains equal to V^* (see Fig. 8.3a) but the currents go on rising and the collector potential dropping off¹.

¹ In Fig. 8.3 and in other figures, we use the method of load lines to determine the quiescent (Q) points. The approach comes to the following. Lay off the given voltage (E_b or E_c here) on the y -axis and from this voltage point construct an I - V curve for the load (R_b or R_c in the given case). The point of intersection of both curves gives the current and voltages in question.

At point 2 corresponding to current I_{b2} , the collector potential V_c becomes equal to V^* and thus the voltage across the collector junction, $V_{cb} = V_c - V_b$, goes to zero. As the current rises still further, the voltage V_{cb} becomes negative, that is, the collector junction becomes forward biased, and the transistor operates in the double injection mode. But so long as the forward bias at the collector junction remains lower than the turn-off voltage (0.6 V, see p. 89), the collector injection is insignificant, so the collector current grows in proportion to the base current. At point 3 the forward

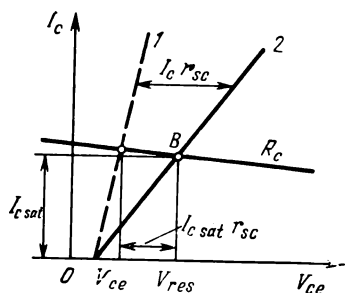


Fig. 8.4. The initial portion of a collector characteristic showing residual voltage components

voltage V_{cb} reaches a value of 0.6 V (the collector potential correspondingly falls to about 0.1 V). Then the collector injection becomes significant and begins to impede the further rise in the collector current, which then remains practically constant.

This maximum collector current is known as the *saturation current*, denoted as $I_{c sat}$ and the double injection mode specific to the on condition of the switch is known as the *saturation mode*. The point B in Fig. 8.3 represents the on (saturated) condition for the switch. The control current and voltage in the on condition are respectively designated as I_b^+ and E_b^+ . The residual voltage across the switch, at point B in Fig. 8.4, contains two components:

$$V_{res} = V_{ce} + I_{c sat} r_{sc} \quad (8.1)$$

where V_{ce} is the voltage difference between the junctions and $I_{c sat} r_{sc}$ is the voltage drop across the series collector resistance (see Fig. 7.6).

The first component is found from Eq. (4.38c), which can be readily reduced to the form

$$V_{ce} = \varphi_T \ln \frac{1}{\alpha_I} \frac{I_b^+ + I_{c sat}/(B_I + 1)}{I_b^+ - I_{c sat}/B_N} \quad (8.2)$$

For example, if $B_N = 100$, $B_I = 1$, and $I_{c sat} = I_b^+$, then $V_{ce} \approx 27$ mV. If $I_{c sat} = 10 I_b^+$, then V_{ce} grows to 65 mV. Note that a decrease in component V_{ce} is due primarily to an *increase in inverse gain* B_I . Thus, if we set $B_I = 3$ in the above examples, then the values of V_{ce} will respectively come to 12 and 40 mV.

The second (ohmic) component varies over a wide range depending on the saturation current and the transistor structure. If $I_{c sat} = 100 \mu A$ and $r_{sc} = 10 \Omega$ (the structure with a buried layer), then $I_{c sat} r_{sc} = 1$ mV, and so the second component is much smaller than

the first component. But if $I_{c\ sat} = 2\text{ mA}$ and $r_{sc} = 100\ \Omega$ (the structure without a buried layer), $I_{c\ sat}r_{sc} = 200\text{ mV}$, and so the second component noticeably exceeds the first. That is why switching ICs always use buried layers. The total residual voltage commonly ranges from 50 to 100 mV.

8.3.2. Saturation criterion. Relying on Fig. 8.3, we can easily derive the expressions for the turn-on base current and collector saturation current:

$$I_b^+ = (E_b^+ - V^*)/R_b \quad (8.3a)$$

$$I_{c\ sat} = (E_c - V_{res})/R_c \approx \frac{E_c}{R_c} \quad (8.3b)$$

Since the voltage V^* is practically constant, it is safe to assume that both currents are functions of **external** parameters such as E_b , R_b , E_c and R_c . In other words, in calculating a saturated switch, we can assume I_b^+ and $I_{c\ sat}$ to be **specified, independent** variables and the voltages to be the functions of the currents.

The formal criterion of switch saturation (that is, switch operation in the mode of double injection) is the condition at which the collector junction is forward biased. In Ebers-Moll formulas (4.32) forward voltages are positive in sign. The saturation criterion can thus be of the form: $V_c > 0$. But this criterion is inconvenient to deal with where the specified variable is **current**. Instead, it is suitable to employ the **current** saturation criterion:

$$BI_b^+ > I_{c\ sat} \quad (8.4)$$

where B is the common-emitter current gain for a transistor in the normal operation region.

The criterion (8.4) is easy to derive from (4.34) if we express V_c in terms of currents, substituting $I_e = I_c + I_b$ and setting $V_c > 0$.

The inequality (8.4) must be strong enough so that the inevitable changes in the quantities entering this criterion should not cause the switch to go out of saturation, otherwise this would lead to a sharp rise in residual voltage.

The relation between the quantities entering the inequality (8.4) is described by a special parameter known as the *degree of saturation*

$$S = BI_b^+/I_{c\ sat} \quad (8.5)$$

The quantity S is equal to unity at the edge of the active region; at the zero collector current, S extends to infinity. When the base and collector currents are equal, $S = B$. In Fig. 8.3b, the degree of saturation approaches 2.

8.3.3. Parallel connection of switches. It is usual practice to use one voltage source E_b to control a few switches, in which case the

emitter junctions of these switches become connected in parallel. Fig. 8.5a shows the circuit of two parallel-connected switches. Obviously, the total current I_b is distributed (divided) between the bases:

$$I_b = I_{b1} + I_{b2}$$

If the transistors and their operating conditions are identical, the total current is equally divided between the bases:

$$I_{b1} = I_{b2} = 1/2 I_b$$

If the input I - V characteristics are nonidentical, the distribution of current I_b at the equal voltages $V_b' = V^*$ can be rather nonuniform (see currents I_{b1}' and I_{b2}' in Fig. 8.5b).

For an analytic estimate of this nonuniformity, let us use Eq. (4.38a). For the switch under analysis, the voltage V_{eb} is the

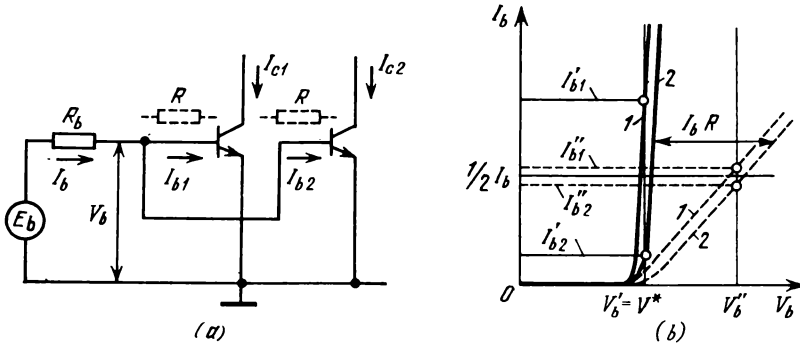


Fig. 8.5. Parallel connection of switches
(a) circuit; (b) current distribution

input voltage V_b . If we pass from the factor α_I to B_I in (4.38a), take into account (4.33), and set $\alpha_N \approx 1$ for simplicity, the expression for V_b assumes the form

$$V_b = \varphi_T \ln \frac{I_c + (B_I + 1) I_b}{B_I I_{c0}} \quad (8.6)$$

This is the expression for the input I - V characteristic of the switch with parameters I_c , B_I and I_{c0} .

The parameters I_{c0} and B_I in integrated circuits show a comparatively small spread because integrated transistors lie close to each other on the chip. As for the currents I_c , these can differ substantially in complex switching circuits. From Eq. (8.6) it is clear that at the same values of V_b , the transistor with a higher value of I_c will exhibit a smaller degree of saturation. With a rather large difference in currents I_c , the transistor with a higher current can fail to reach saturation at all, which disturbs the operation of the switch.

It is easy to see that a substantial difference between the base currents in the parallel-connected switches stems from a steep rise of the I - V curve in the vicinity of voltage V^* . It is thus possible to equalize currents I_b by decreasing the steepness of the I - V curve. For this, one should connect series resistors R of identical values to the base circuits of all the transistors, as shown by dash lines in Fig. 8.5a.

The resultant I - V curves are plotted by points. To do this, set the current I_b and add up the corresponding voltages V_b and $I_b R$ to obtain one of the points of an I - V curve. Next, draw a smooth line through all the points obtained in the same way. The resultant curves appear as shown in Fig. 8.5b by dash lines. As seen from the figure, these are practically straight lines originating from the point V^* and showing a slope corresponding to the resistance R . For these I - V curves, nearly identical currents I_{b1}' and I_{b2}' correspond to the voltage V_b'' .

Note that the base resistance r_b plays the same part as the external resistor R . But the typical values of r_b , ranging from 100 to 200 Ω , are insufficient to smooth out the I - V curve to the desired degree. The sum $r_b + R$ may formally be regarded as an equivalent base resistance. The transistor with such a large resistance exhibits an increased forward voltage across the emitter junction, $V_e = V^* + I_b R$. Assuming $I_b = 1$ mA and $R = 0.5$ k Ω , then V_e averages 1.2 V.

8.3.4. Series switching circuit. Individual switches find use mainly in analog circuits. Characteristic of digital circuits is a joint operation of a few switches forming what is known as a *switching network*. In such a network, the preceding switch controls the next switch and this in turn controls the operation of the subsequent switch.

Consider the series network of switches illustrated in Fig. 8.6. We shall disregard for the time being the switch $T4$ shown by a dash line. If the transistor $T1$ is on in saturation, the potential V_{c1} and the equal potential V_{b2} come close to zero, and hence the transistor $T2$ is in the off condition. A current then flows into the base of $T3$ from the source E_c via R_c , so the transistor $T3$ is on. *The series network thus features alternately open and closed switches.*

The equivalent circuits for the transistors in saturation and cutoff appear in Fig. 8.7. These equivalent circuits permit us to determine the operating conditions for the base circuits of controlled transistors

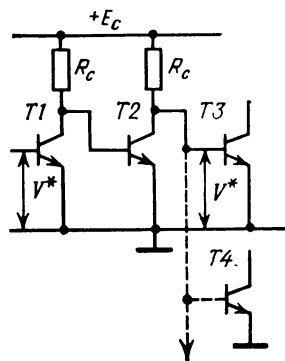


Fig. 8.6. Switching network

Comparing the circuit of Fig. 8.7a with the general circuit of Fig. 8.2, we can write the parameters of the base circuit in the off switch:

$$E_b^- = V_{res} \approx 0, \quad R_b \approx 0 \quad (8.7a)$$

where E_b^- is the "cutoff" voltage. Comparing the circuit of Fig. 8.7b

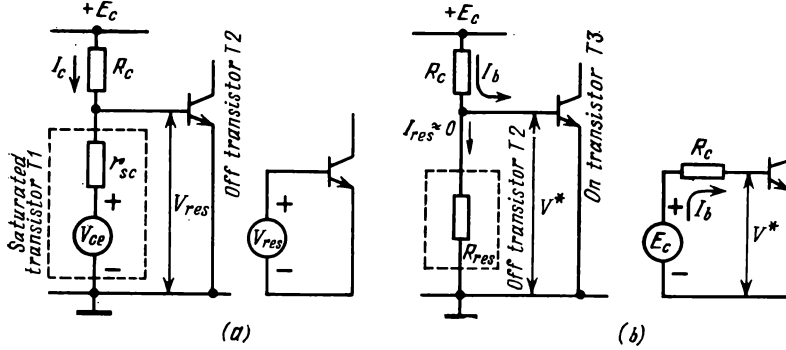


Fig. 8.7. Switch circuit models
(a) off condition; (b) on condition in saturation

with the general circuit of Fig. 8.2 gives the parameters of the base circuit in the on switch:

$$E_b^+ \approx E_c, \quad R_b \approx R_c \quad (8.7b)$$

Substituting the values of E_b^+ and R_b from Eq. (8.7b) into (8.3a) gives the turn-on base current:

$$I_b^+ = (E_c - V^*)/R_c \quad (8.8)$$

The saturation collector current is expressed by the same formula (8.3b) as for an "isolated" transistor switch:

$$I_{c \text{ sat}} = (E_c - V_{res})/R_c \approx E_c/R_c \quad (8.9)$$

As seen, the collector and base currents in the series network are almost equal.

Substituting (8.8) and (8.9) into the saturation condition (8.4), we set the limit on the current gain:

$$B > E_c/(E_c - V^*) \quad (8.10)$$

Even at minimum supply voltage, $E_c = 1$ to 1.2 V, we get $B > 2.4$ to 3.3 , which is easy to achieve under common operating conditions.

The degree of switch saturation in the series circuit, according to (8.5), will take on the form

$$S = B (E_c - V^*)/E_c \quad (8.11)$$

At supply voltages, typically 3 to 5 V, the degree of saturation may reach 50 to 100 and more.

Let us point out in conclusion that the voltage range typical of a switch operating in the series network is narrower than for an isolated switch. Indeed, the cutoff voltage in an isolated switch is $V_c = E_c$ (see point *A* in Fig. 8.3*b*), whereas for a switch operating in the series network (*T2* in Fig. 8.6), $V_c = V^*$. This large difference is due to the fact that the collector of the off transistor is connected to the base of the next, on transistor.

8.3.5. Load capacity of a switch. Switching circuits typically comprise a combination of series- and parallel-connected switches; namely, in a series network each transistor can control not one but a few parallel-connected switches. Thus, as shown in Fig. 8.6 by dash lines, the transistor *T2* can control not only the switch *T3* but at the same time the switch *T4*; in the general case, *T2* can ensure control over a number of switches.

The load capacity of a switch is its ability to control the operation of a number of parallel-connected switches. Denote this number by *n*.

Suppose the total turn-on current of Eq. (8.8) is equally divided among *n* bases. Then, in each of the parallel-connected switches

$$I_{b1}^+ = \frac{1}{n} \frac{E_c - V^*}{R_c^*}$$

The current I_{b1}^+ must satisfy the saturation condition (8.4), where the collector current $I_{c\ sat}$ is defined by (8.9) as before. From the condition (8.4) we can readily determine the principal limit on the load capacity:

$$n < \frac{E_c - V^*}{E_c} B \quad (8.12a)$$

In reality, the limit must be more rigid because it is necessary to ensure not just saturation, that is, the condition $S > 1$, but a definite minimum degree of saturation, S_{min} . Then, considering (8.5), the inequality for the load capacity limit will take the form

$$n < \frac{B}{S_{min}} \frac{E_c - V^*}{E_c} \quad (8.12b)$$

Assuming $E_c = 3$ V, $B = 100$, and $S_{min} = 4$, we have $n \leq 18$. Setting the value of S_{min} , we should consider, in particular, the spread in base currents due to nonidentical input I - V curves (see Subsec. 8.3.3).

8.4. Transients in a Simple Bipolar Switch

Transients occur because of a steep-like changes in the input signal. Fig. 8.8 shows the time relationships for currents and voltages in a transistor switch.

The output pulses i_c and v_c are seen to be shifted with respect to the step input e_b ; the leading edge (front) and the trailing edge (tail) of the output pulses have finite durations. It is thus usual to differentiate between the *delay time* t_d for the leading edge and the *delay*

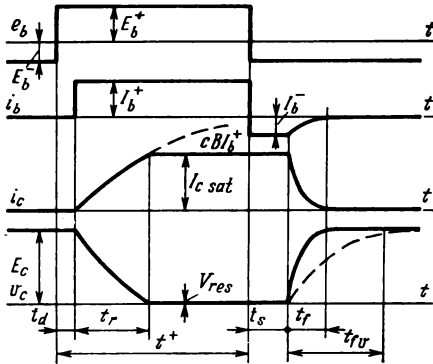


Fig. 8.8. Transients in a bipolar switch

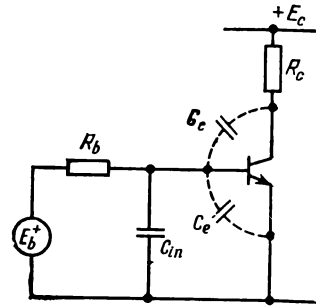


Fig. 8.9. The circuit model for a switch at an instant of leading edge delay

time, or *storage time* t_s , for the trailing edge, and also the *rise time* t_r for the front and the *fall time* t_f for the tail. The relation between the increments of the collector voltage and the increments of the collector current has the form

$$v_c = -i_c R_c$$

8.4.1. Turn-on delay. This first stage of the transient results from **charging of the input capacitance** of the off transistor (Fig. 8.9). The charge begins to build up as the control voltage changes stepwise from E_b^- to E_b^+ . The process of charging is given by

$$v_b(t) = E_b^+ (1 - e^{-t/\tau_c}) - E_b^- e^{-t/\tau_c}$$

where $\tau_c = C_{in} R_b$ is the time constant of charging.

As the voltage V_b grows and becomes equal to V^* , the emitter junction of the transistor starts conducting, so the stage of charging terminates. We thus can find the delay time t_d by setting $v_b(t) =$

$= V^*$; its expression has the form

$$t_d = \tau_c \ln \frac{E_b^+ + E_b^-}{E_b^+ - V^*} \quad (8.13)$$

Assuming $E_b^- = 0$ according to (8.7a) and $E_b^+ = 3V$, we get $t_d \approx 0.25\tau_c$.

The input capacitance is usually taken equal to the sum of the barrier capacitances of the emitter and collector junctions:

$$C_{in} = C_e + C_c \quad (8.14)$$

Setting $C_{in} = 2$ pF and $R_b = 2$ k Ω , we find $\tau_c = 4$ ns. For the above values of E_b^- and E_b^+ , we have $t_d \approx 1$ ns.

8.4.2. Front forming. A rise in the collector current and fall in the collector voltage at the second stage of the transient takes place at the **base drive current** I_b^+ given by Eq. (8.3a). In the analysis, therefore, we should use the equivalent time constant of Eq. (4.67). Write this expression in the form

$$\tau_{oe} = \tau + (B + 1) C_c R_c \quad (8.15)$$

The lifetime can be taken to equal 100 ns in the absence of gold as a dopant, and 10 ns if gold is present. The time constant of collector capacitance averages 100 ns at typical values of $C_c = 0.5$ pF, $B = 100$, and $R_c = 2$ k Ω . This time constant is thus always rather large and plays a decisive part in transistors doped with gold. In the examples that follow we assume that the average value of τ_{oe} is 150 ns.

Substituting the quantity B (s) from (4.58) into the relation $I_c = BI_b$, setting $I_b = I_b^+$, $\tau_B = \tau_{oe}$, and passing to the original function we obtain

$$i_c(t) = BI_b^+ (1 - e^{-t/\tau_{oe}}) \quad (8.16)$$

The asymptotic value of current at t going into infinity is $I_c(\infty) = BI_b^+$ (see Fig. 8.8). But this value cannot be reached because at the moment t_r the current assumes the value $I_{c\text{ sat}}$, corresponding to the saturation region, and so the initial relation $I_c = BI_b$ is no longer valid. The rise time is easy to find from (8.16) substituting $i_c(t) = I_{c\text{ sat}}$:

$$t_r = \tau_{oe} \ln BI_b^+ / (BI_b^+ - I_{c\text{ sat}}) \quad (8.17)$$

Let us assume $B = 100$. If $I_b^+ = 0.1 I_{c\text{ sat}}$, then $t_r = 0.1 \tau_{oe}$, and if $I_b^+ = I_{c\text{ sat}}$, then $t_r = 0.01 \tau_{oe}$. Setting $\tau_{oe} = 150$ ns, we get $t_r \approx 15$ ns for the first case and $t_r = 1.5$ ns for the second.

From Eq. (8.17) it follows that *as the turn-on current I_g^+ grows, the rise time decreases.*

8.4.3. Charge accumulation. With the transistor driven into saturation, any noticeable **external** changes in the switch circuit do not occur. But the charge continues to grow not only in the base but also in the collector layer (see Fig. 4.9).

At the beginning of this stage, when the transistor operates at the edge of the active region, the charge is defined by Eq. (4.10b). Substituting $I_e = I_{c\ sat} + I_b^+$ and considering (4.46), we can write the *boundary charge* in the form

$$Q_{bd} = (I_{c\ sat} + I_b^+) t_{tr} \quad (8.18)$$

At the end of this stage, the **stationary** charge is determined by Eq. (4.12). Substituting $I_b = I_b^+$, we get

$$Q^+ = I_b^+ \bar{\tau} \quad (8.19)$$

where $\bar{\tau}$ is the mean lifetime of carriers in the base and collector layers.

The relation between charges Q_{bd} and Q^+ depends on the relation between currents $I_{c\ sat}$ and I_b^+ , that is, on the degree of saturation. In most switches, the currents $I_{c\ sat}$ and I_b^+ are comparable in magnitude, whereas the parameters t_{tr} and $\bar{\tau}$ are sharply different; $t_{tr} \ll \bar{\tau}$. In practice, therefore, Q_{bd} is typically much smaller than Q^+ .

Since the currents do not change at this stage, the charge grows only because of the thermal generation of carriers, and hence the rate of charge accumulation is a function of the lifetime. The process of charge buildup is exponential in character, the corresponding expression being

$$Q(t) = Q(0) e^{-t/\tau} + Q(\infty) (1 - e^{-t/\tau}) \quad (8.20)$$

where $Q(0)$ and $Q(\infty)$ are the initial and the steady-state values of charge respectively.

Considering the inequality $Q_{bd} \ll Q^+$, we may neglect the charge $Q(0)$ in Eq. (8.20). Further, we may assume that the process of charge accumulation comes to an end at the 95% level of $Q(\infty)$. From Eq. (8.20) we then find the charge *accumulation time*

$$t_{ac} = 3\bar{\tau} \quad (8.21)$$

If $\bar{\tau} = 30$ ns, $t_{ac} \approx 90$ ns.

In order for the charge accumulation to have time to come to completion, the length t^+ of the turn-on pulse (see Fig. 8.8) should be larger than the accumulation time t_{ac} . Otherwise, at the moment of reverse switching the stored charge will be smaller than the steady value $Q(\infty)$; namely, if we neglect the charge $Q(0)$, as done before, then at the moment t^+ the charge will be equal to

$$Q(t^+) \approx I_b^+ \bar{\tau} (1 - e^{-t^+/\bar{\tau}}) \quad (8.22)$$

In the case of short pulses, this charge can be equal to fractions of $Q(\infty)$.

8.4.4. Turn-off delay. *The charge stored in the layers and junctions of a transistor cannot change instantly, and so the voltages at the emitter and collector junctions cannot change instantly too. So at the moment of switching the voltage E_b from E_b^+ to E_b^- (to zero, in particular), both pn junctions are under **forward** bias, close to V^* , and the collector current remains equal to $I_{c\ sat}$. The base current, however, takes on the value*

$$I_b^- = (E_b^- - V^*)/R_b \quad (8.23)$$

If the switch operates in a series network, the cutoff condition ensues as the preceding switch comes to saturation. The cutoff current I_b^- can then be found from the equivalent circuit of Fig. 8.10. In this circuit, the preceding saturated transistor is represented by the resistance r_{sc} and voltage V_{ce} (cf. Fig. 8.7a); as for the transistor going to cutoff, the base resistance r_b is shown separately. Disregarding in a first approximation the current I_b^+ , voltage V_{ce} , and resistance r_{sc} , we find the **initial** value of cutoff current

$$I_b^- = -V^*/r_b \quad (8.24)$$

If we set $r_b = 100\ \Omega$, then $I_b^- \approx -7\text{ mA}$; this current can be a few times the turn-on current I_b^+ , particularly if the switch handles currents in the microampere range. If we include the quantities disregarded above, the current I_b^- will be a little smaller.

A jump of base current from I_b^+ to I_b^- leads to a smooth decrease, or what is called *dissipation* of the charge from Q^+ to Q_{bd} values given by (8.19) and (8.18).

The charge dissipation takes place under the same conditions as charge accumulation, namely, at invariable external currents. The rate of charge dissipation is therefore determined by the same time constant $\bar{\tau}$, and the charge dissipation equation coincides with charge accumulation equation (8.20). However, the initial and steady-state values of charge will be different.

The initial charge at the stage of its dissipation will be equal to the finite charge at the stage of accumulation. At a sufficiently long turn-on pulse, this charge is described by Eq. (8.19), for which reason we assume

$$Q(0) = I_b^+ \bar{\tau} \quad (8.25a)$$

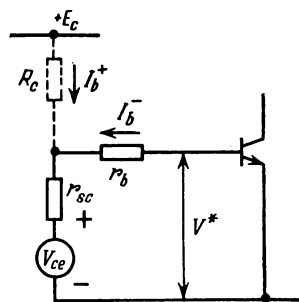


Fig. 8.10. The circuit model of a switch at an instant of charge dissipation with the preceding switch in saturation

The steady-state (asymptotic) charge is determined as usual by the base current, by the cutoff current I_b^- in the given case:

$$Q(\infty) = I_b^- \tau \quad (8.25b)$$

Since $I_b^- < 0$, the quantity $Q(\infty)$ becomes negative. This means that for the case under study this quantity should be regarded not as a real charge but only as its asymptotic value.

At the end of charge dissipation, the concentration of excess carriers at the collector-base boundary drops to zero, which restores the reverse bias at the collector junction. It is only after this moment that the collector current can begin to decay and the trailing edge to shape up.

The *trailing edge delay time*, or the *storage time* t_s , is the time interval during which the charge decreases from its initial value $Q(0)$ to its residual value Q_{res} corresponding to the beginning of reverse biasing of the collector junction¹.

The residual charge is generally much smaller than the boundary charge given by Eq. (8.18), and the latter, in turn, is much smaller than the charge stored up in the saturation condition (see p. 277). As a first approximation, therefore, we can ignore the residual charge and determine the storage time from (8.20) setting $Q(t) = 0$.

Then, taking into account Eq. (8.25), we have

$$t_s = \tau \ln \left(1 + \frac{I_b^+}{|I_b^-|} \right) \quad (8.26)$$

where $|I_b^-|$ is the absolute value of cutoff current. Thus, if the ratio between the currents ranges from 0.2 to 5.0, the storage time is 0.2τ to 1.6τ , or 6 to 50 ns at $\tau = 30$ ns.

From Eq. (8.26) it is clear that *the storage time (tail delay time) shortens with decreased turn-on current I_b^+ and increased cutoff current I_b^-* . Since a small value of current I_b^+ leads to an increase in the rise time of Eq. (8.17), it is primarily desirable to **raise the cutoff current I_b^-** . This approach finds wide use in designing transistor switches.

In the preceding analysis we have implied without reservation that the cutoff current I_b^- is rather small as against the saturation current $I_{c sat}$. Only this assumption allows us to suppose that during charge dissipation the distribution curves of Fig. 4.9 "return" to the initial boundary curves "in the same sequence" as during charge storage, but of course in the reverse direction. For a diffusion transistor, the distribution curve shifts downward parallel to itself; in

¹ The term dissipation time adopted in Soviet literature for t_s seems to be more adequate for the physical process associated with this quantity (I.P. Stepanenko).

a drift transistor, the curve changes in shape, passing from the curve given in Fig. 4.9b to the curve of type 1-4 as shown in Fig. 4.6.

If the condition $I_b < I_{c\text{ sat}}$ does not hold, as is practically the case, the distribution curves for charge dissipation differ in shape from the distribution curves during charge storage. As a result, the curve corresponding to the boundary charge changes in principle: it falls down to zero at the emitter-base boundary (see Figs. 4.7 and 4.8) rather than at the collector-base boundary (see Figs. 4.5, 4.6). This means that the emitter junction becomes reverse biased earlier than the collector junction. For this condition the term *inverse*, or *emitter, charge dissipation* is adopted in contrast to the term *normal*, or *collector, charge dissipation* taking place when the reverse bias first sets in at the collector junction.

It is easy to guess that with emitter dissipation the delay in tail forming will be longer than with collector dissipation. Analysis gives the following expression for the inverse storage time:

$$t_{sI} = \bar{\tau} \ln \frac{1 + I_b^+ / |I_b^-|}{1 - I_{cs} / B_I |I_b^-|} \quad (8.27)$$

where B_I is the inverse common-emitter current gain (see Subsec. 4.4.4). Assuming, $B_I = 1$ and $I_b^+ = |I_b^-| = 2I_{c\text{ sat}}$, we get $t_{sI} \approx 1.4 \bar{\tau}$, whereas in the case of normal charge dissipation $t_s \approx 0.7\bar{\tau}$.

8.4.5. Tail forming. After the stage of charge dissipation is completed, the last stage of the transient commences to turn the transistor off. This stage is most difficult to analyze quantitatively because the residual charge is rather small and the shape of its distribution in the base is complex.

Consider first the simplest case where the cutoff current I_b^- is so small that at the beginning of switch-off the carrier distribution corresponds to the normal active mode of operation (Fig. 8.11a). The initial charge $Q(0)$ is then equal to the boundary charge Q_{bd} described by Eq. (8.18). A further decay of the charge proceeds with the same time constant as for the decay of collector current, that is, at τ_{oe} given by (8.15). The asymptotic value of the charge remains the same as it is at the stage of charge dissipation: $Q(\infty) = Q^-$. The quantity Q^- is determined by Eq. (8.25b). The formation of the trailing edge terminates when the collector current, along with the charge, becomes zero.

Substituting $Q(t) = 0$ in Eq. (8.20), replacing $\bar{\tau}$ by τ_{oe} in the exponent, and using the above expressions for $Q(0)$ and $Q(\infty)$, we find the pulse fall time (trailing edge duration):

$$t_f = \tau_{oe} \ln \left(1 + \frac{I_{c\text{ sat}} t_{tr}}{|I_b^-| \bar{\tau}} \right) \quad (8.28)$$

The initial conditions set up in deriving Eq. (8.28) are typical for **rather small** cutoff currents, much smaller than the saturation current. Such a situation rarely arises in practice. The cutoff current is more often comparable to or in excess of the saturation current.

Figure 8.11b shows the curves of initial carrier distribution at a sufficiently large cutoff current, which is merely a few times smaller than the current $I_{c\ sat}$. For comparison, Fig. 8.11b also illustrates in dash lines the distribution curves shown in Fig. 8.11a. As seen, the areas under the curves are noticeably smaller than when the switch operates in the active region. This supports the conclusion that the

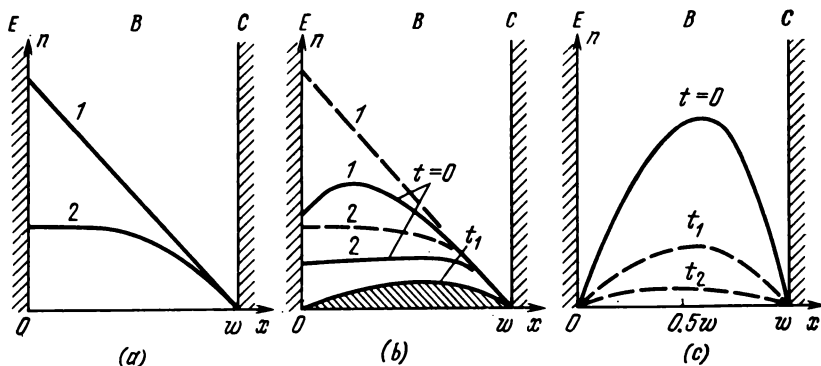


Fig. 8.11. Carrier concentration distribution in the base during reverse biasing at small (a), moderate (b), and large (c) cutoff currents

1—diffusion transistor; 2—drift transistor

residual charge at the end of carrier dissipation is smaller than the boundary charge (see p. 278). The fall time will then be shorter than if we calculate it with Eq. (8.28). But the main feature of the given case consists in a **change of the structure** of the trailing edge rather than in the above quantitative difference.

The thing is that at a certain moment t_1 the carrier concentration at the emitter-base boundary also goes to zero. As this takes place, both the collector and emitter junctions operate under **reverse bias**. Formally, we should regard this region of transistor operation as a **cutoff region**. However, in distinction to the “traditional” cutoff condition, at which the excess charge in the base is zero and junction currents are negligible, in the case under consideration a certain **residual** base charge is still present (corresponding to the hatched area). Thus despite the reverse bias on the junctions, they have quite finite values of current. This condition of transistor operation is known as **dynamic cutoff**.

At dynamic cutoff, all the three currents of a transistor drop to zero with a cutoff time constant τ_{cut} , whose expression is given below. The time constant τ_{cut} is significantly smaller than τ_{oe} which

determines the pulse tail forming process up to the moment t_1 . That is why the end part of the pulse tail will be steeper than its initial portion.

The cutoff time constant can be roughly estimated proceeding from the fact that the excess carriers in the middle portion of the base have to travel a distance $1/2w$ in order to leave the base through one of the junctions (see Fig. 8.11c). The corresponding time according to Eq. (4.46) will be one-fourth the transit time for carriers crossing the **entire** base. Therefore, using Eqs. (4.47a) and (4.66), we can write the expression for cutoff time constant in the form

$$\tau_{cut} = 0.25t_{tr} + C_c R_c \quad (8.29)$$

The first term is of no consequence in most cases, and so the time constant τ_{cut} is primarily determined by the time constant of collector capacitance. For example, if $t_{tr} = 0.2$ ns, $C_c = 0.5$ pF, and $R_c = 1$ k Ω , then $\tau_{cut} \approx C_c R_c = 0.5$ ns. The fall time at a 10% level of the initial current I_{cs} is expressed by

$$t_f = 2.3\tau_{cut} \quad (8.30a)$$

At $\tau_{cut} = 0.5$ ns, we have $t_f \approx 1.1$ ns.

The preceding analysis dealt with the normal (collector) charge dissipation. In the case of emitter charge dissipation, the transient shows some distinguishing features. At the moment $t = 0$, when the carrier concentration at the emitter-base boundary comes to an equilibrium, the concentration at the collector-base boundary does not yet exceed the equilibrium value (dash line $t = 0$ in Fig. 8.11c). This means that the collector junction stays at the forward bias for a certain length of time, that is, the voltage across this junction remains invariable and equal to V^* .

At the same time the reverse voltage grows at the emitter junction, so the base potential becomes more and more negative. A negative increment in potential V_b is fully passed to the collector through the forward-biased collector junction. A negative voltage surge and positive current surge then appear at the collector. These surges, shown by points in Fig. 8.8, go to zero when the carrier concentration at the collector-base boundary drops to an equilibrium value. After this, the trailing edge begins to form with a time constant τ_{cut} , as mentioned above.

8.4.6. Effect of load capacitance. If the switch operates into a capacitive load (the input of the next off-transistor, see Fig. 8.9), then the leading and trailing edge times of a current pulse can differ substantially from those of a voltage pulse. For example, during tail formation, the current decays as usual with a small time constant τ_{cut} , whereas the voltage rises at a much greater time constant $C_l R_c$, where C_l is the load capacitance (see the dash curve in Fig. 8.8).

If C_i is much larger than C_c , which is often the case, the current pulse trailing edge may practically be considered to be a vertical, but the voltage pulse trailing edge (see Fig. 8.8) has a finite length expressed as

$$t_{fv} \approx 2.3C_iR_c \quad (8.30b)$$

Similar differences between the rise times of current and voltage take place in turning the switch on if the load capacitance exceeds the collector capacitance. The current pulse leading edge here may practically be regarded to rise vertically, but the voltage pulse leading edge rises in accordance with Eq. (8.30b).

8.5. Schottky-Barrier Transistor Switch

One of the basic problems involved in improving the transient response of transistor switches is to shorten the storage time, that is, the time taken for dissipation of the excess charge. For this, as seen from (8.26), it is necessary to decrease the turning-on current I_b^+ , that is, the degree of saturation, S . But this tends to lengthen the rise time as follows from (8.17). Besides, in practical cases, the degree of saturation must be in excess of a minimum value S_{\min} (see Subsec. 8.3.5), otherwise the slightest decrease in the gain B or current I_b^+ can drive

the transistor into the active mode of operation, which entails an increase in the residual voltage on the switch.

A generally accepted method for preventing a transistor from going into saturation and, at the same time, escaping the above complications is to use **nonlinear feedback** in the switch. This method suggested by B. N. Kononov back in 1955 consists in connecting a clamping diode between the collector and base of the transistor (Fig. 8.12).

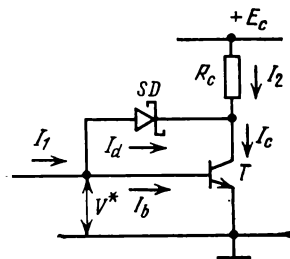


Fig. 8.12. Switch having nonlinear feedback (Schottky-barrier transistor)

When the transistor is off or operates in the active region, the potential on the collector is positive with respect to the base. The diode is then reverse biased and does not affect the operation of the switch. If during formation of the leading edge the collector potential *with respect to the base* traverses zero and becomes negative, the diode switches on and remains at a forward voltage V_d^+ . If this voltage is lower than 0.5 V, which is typical of Schottky diodes (see Subsec. 3.4.1), the **collector junction** is practically cut off. This excludes collector injection and thus the excess charge storage specific to the saturation mode. So at switch-off, there will be no such a stage as excess charge dissipation and turn-off delay. The considered com-

bination of a transistor and Schottky diode received the name *Schottky barrier transistor*.

It is easy to see that saturation in the Schottky-barrier transistor does not occur because the diode forward voltage V_d^* is lower than the forward voltage V^* at the silicon pn junction. If the switch had a conventional diode (pn junction) in place of the Schottky diode, it would be necessary to decrease the forward voltage at the diode artificially (by a circuit design approach), connecting the diode in series with an emf $e = -0.2$ or -3 V. That was precisely the approach until the advent of Schottky diodes in the 1960s.

The residual voltage on a Schottky-barrier switch is somewhat greater than that on a conventional transistor switch; namely,

$$V_{res} = V^* - V_d^* = 0.2 \text{ or } 0.3 \text{ V}$$

But this shortcoming is compensated for by a faster switching speed since the transistor works all the time in the active region.

It should be pointed out that despite the elimination of saturation, the Schottky-barrier switch is little sensitive to changes in the gain B and turn-on current, since the residual voltage weakly depends on these quantities and hence the current I_2 retains its value as defined by Eq. (8.3b).

The delay time t_d and rise time t_r are the same as in the saturated transistor. But the switch-off process takes another course.

When the control current I_1 assumes the value I_1^- , the current I_d remains stable at the first moment, but the current I_b changes by the same value as the current I_1 :

$$\Delta I_b = \Delta I_1 = I_1^- - I_1^+ < 0$$

Since the transistor operates in the active region, the surge ΔI_b causes the collector current to fall off with a time constant τ_{oe} . The increments ΔI_c then fully flow through the diode D and reduce its current I_d :

$$i_d(t) = I_d - B |\Delta I_b| (1 - e^{-t/\tau_{oe}}) \quad (8.31)$$

where $|\Delta I_b|$ is the absolute value of the base current increment.

The current I_2 begins to decrease only when the diode switches off. Assuming that the left-hand side of (8.31) is equal to zero, we can readily find the storage time t_s .

If, besides, we assume $1/B \ll 1 - I_1^-/I_1^+$, which is commonly the case in practice, then, expanding the logarithm into a series, we get

$$t_s \approx \frac{\tau_{oe}}{B} \frac{1}{1 + |I_1^-|/I_1^+} \quad (8.32)$$

Thus, setting $|I_1^-|/I_1^+ = 1$, we have $t_s = \tau_{oe}/2B$. In modern transistors this quantity rarely exceeds 0.5 ns.

8.6. Current Switch

The current switch is a symmetric circuit (Fig. 8.13) in which the specified current I_0 passes through one of the branches depending on the voltage V_b at the controlling input. The potential E at the other input is kept constant.

So the first special feature of the current switch is that the control parameter here is **voltage** rather than current as is the case in the simplest switch. The second feature is that the *on* transistors operate in the **nonsaturated** (active) region, which offers a faster switching speed owing to the elimination of the storage time.

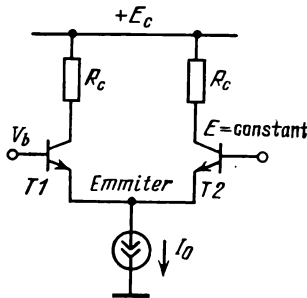


Fig. 8.13. Current switch

8.6.1. Static operation. Let us first set $V_b = E$. In this condition, both transistors are conducting and in both emitter currents are $0.5 I_0$. The potential on the emitters is lower than the potential E by a value V^* : $V_e = E - V^*$.

Decrease now the potential V_b by a value $\delta \geq 0.1$ V. Since the potential V_e remains constant, the voltage V_{be1} will drop by a value δ . The current in transistor $T1$ will then decrease by a factor of a few tens (see p. 89). So the input pulse $V_b \leq E - \delta$ will switch off the transistor $T1$, while the total current I_0 will flow through the transistor $T2$. We shall call the quantity

$$E_b^- = E - \delta \quad (8.33a)$$

the *cutoff potential*.

If, on the contrary, we shall raise the potential V_b by a value δ , the potential on the emitters will increase by the same amount and the voltage V_{be2} will drop accordingly; the current through the transistor $T2$ will then sharply decrease. So at the input pulse $V_b \geq E + \delta$ the transistor $T2$ can be considered nonconducting, and the total current I_0 passes through the transistor $T1$. We shall call the quantity

$$E_b^+ = E + \delta \quad (8.33b)$$

the *turn-on potential*.

Thus, the variation of potential, $\Delta V_b = \pm \delta$ about the mean value of E ensures switching of the current I_0 from one transistor to the other. The relations between the turn-on and cutoff potentials are as follows:

$$E_b^+ - E_b^- = 2\delta \quad (8.34a)$$

$$1/2 (E_b^+ + E_b^-) = E \quad (8.34b)$$

ted via additional matching circuits known as *level-shifting circuits* (*dc level shifters*).

The simplest method of level shifting is to insert an emf e between the neighbour switches as shown in Fig. 8.14. In this case, with a transistor $T1$ in the n th switch being in the off state, the base potential of a transistor $T1'$ in the $(n + 1)$ th switch will be

$$(V'_{b1})^+ = E_c - e \quad (8.39a)$$

This value must exceed E_b^+ to make the transistor $T1'$ be on. There is no difficulty in meeting the condition $(V'_{b1})^+ > E_b^+$.

If, on the other hand, the transistor $T1$ in the n th switch is off, then the base potential of transistor $T1'$ becomes

$$(V'_{b1})^- = E_c - \alpha I_0 R_c - e \quad (8.39b)$$

This value must be smaller than E_b^- to keep $T1'$ in the off condition.

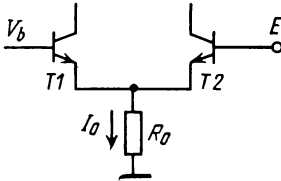


Fig. 8.15. Current switch with a resistor forming a current source I_0

This condition places a certain limit on the emf e . Indeed, let us substitute $E_c - \alpha I_0 R_c$ from (8.37b) into (8.39b). We have

$$(V'_{b1})^- = E + \delta - e$$

Substituting the found quantity $(V'_{b1})^-$ and E_b^- from Eq. (8.34a) into the inequality $(V'_{b1})^- \leq E_b^-$, it is easy to determine the above mentioned limit:

$$e \geq 2\delta \quad (8.40)$$

The practical method of level shifting, that is, realization of emf e , is discussed in Subsec. 10.2.3 (see Fig. 10.3).

The current source I_0 can be accomplished by various methods (see Sec. 9.11). The simplest and historically first method uses the resistor R_0 (Fig. 8.15). If the transistor $T2$ is on, the current I_{02} is determined by the relation.

$$I_{02} = V_e / R_0 = (E - V^*) / R_0$$

If the transistor $T1$ is on, the current I_{01} has a somewhat larger value:

$$I_{01} = V_e / R_0 = (E_b^+ - V^*) / R_0$$

Substituting Eq. (8.37b), we readily find

$$I_{01} = I_{02} + (\delta / R_0)$$

So, the current I_0 does not remain constant in switching, but changes by δ/R_0 . To make this variation negligible, the following condition must hold

$$\frac{\delta}{R_0} \ll \frac{E - V^*}{R_0} \text{ or } E - V^* \gg \delta$$

Thus, if $\delta = 0.1$ V, the voltage E must exceed 1.7 V.

8.6.3. Transients. We assume that control signals go from the emf source having a zero internal resistance. Such an assumption is generally justifiable for practical circuits.

Let the input of the circuit shown in Fig. 8.13 initially stay at the cutoff voltage E_b^- which switches the transistor $T1$ off. When a signal equal to the turn-on voltage E_b^+ arrives at the input, the first stage of the transient process will be charging of the input capacitance as in the simplest switch.

The analysis similar to that performed in Subsec. 8.4.1 gives the expression

$$t_d = \tau_c \ln 2 \approx 0.7\tau_c \quad (8.41)$$

where time constant $\tau_c = r_b C_{in}$. Setting $r_b = 100 \Omega$ and $C_{in} = 2$ pF, we have $t_d \approx 0.15$ ns. If we take into account the finite internal resistance of the signal source, the delay time will be longer accordingly. Eq. (8.41) is valid for any signals **symmetric** about the potential E .

After biasing the transistor $T1$ to the on state, an invariable current I_0 passes through its emitter junction, while the base potential remains a constant value E_b^+ . These conditions mean that as a matter of fact the *transistor is connected in a common-base (CB) configuration*, though **externally** it appears to be in the CE configuration.

At the first moment when the collector current is still zero, the whole emitter current I_0 flows through the base: $I_b(0) = I_0$. The value of $I_b(0)$ can be by far greater than the steady-state value I_b . As the collector current rises, the base current decreases.

As known, the collector current in the CB circuit changes with an equivalent time constant of Eq. (4.66):

$$\tau_{\alpha oe} = \tau_\alpha + C_c R_c \quad (8.42)$$

Because transistors in current switches operate in the active region, the leading and trailing edges change exponentially. The rise time and the fall time at the 10% to 90% level of ΔI_c , where $\Delta I_c = \alpha I_0$, are the same:

$$t_r = t_f = 2.2 \tau_{\alpha oe} \quad (8.43)$$

Expression (8.42) implies that it is desirable that the second term $C_c R_c$ can be made close to or smaller than the first term. It is thus

expedient to choose the resistance R_c relying on the condition

$$R_c \leq \tau_\alpha / C_c \quad (8.44)$$

For example, if $\tau_\alpha = 0.2$ ns and $C_c = 0.5$ pF, then $R_c \leq 0.4$ k Ω .

The accumulation and dissipation of carriers in the nonsaturated switch do not take place and hence turn-off delay is zero.

8.7. MOS Transistor Switches

As in the case of bipolar transistor switches, the static parameters of MOS transistor switches include a *residual current* (in the off condition) and *residual voltage* (in the on condition). There are three types of MOS transistor switches: MOS switches with a *resistive load*, MOS switches with a *dynamic* (transistor) load, and *complementary* (CMOS) switches. The latter use complementary MOSFETs, one of which is *p*-channel and the other *n*-channel. Consider in turn the static parameters of the above types of switches.

8.7.1. Resistive-load switch. The circuit of such a switch using an *n*-channel transistor (NMOS switch) appears in Fig. 3.16a¹. The

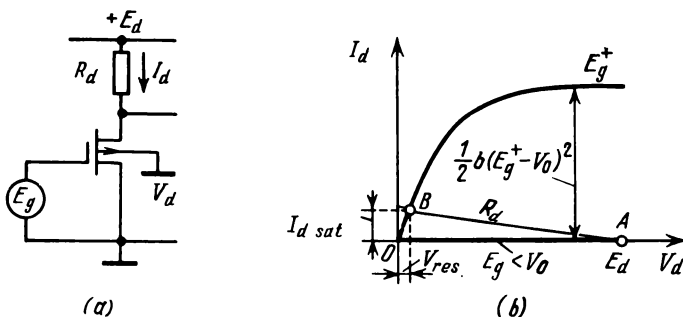


Fig. 8.16. MOS transistor switch with resistive load

(a) circuit; (b) operating points on output characteristic

switch is turned off by applying a gate voltage $E_g^- < V_0$, where V_0 is the threshold voltage (see Sec. 5.2).

In the off transistor, the residual current is the reverse current through the drain *pn* junction because this junction operates under a reverse bias, close to E_d . So the current I_{res} is not more that 10^{-9} to 10^{-10} A, provided the chip has a well polished surface without conducting paths (see Sec. 3.5). On the *I-V* characteristic, the point A

¹ Where the switch uses a *p*-channel MOSFET (PMOS switch), all the voltages in the subsequent analysis should be regarded as absolute values of negative quantities.

corresponds to the off condition of the switch (Fig. 8.16b). At the above values of residual current, we may neglect the voltage drop $I_d R_d$ and assume that the maximum voltage across the off switch is $V_{\max} = E_d$.

A voltage $E_g^+ > V_0$ applied to the gate turns the switch on. This voltage should be large enough to enable the operating point B in Fig. 8.16b to correspond to a minimum voltage possible. As with a bipolar switch, the operating current (saturation current) of this switch in the on condition is determined by the **external** elements of the circuit:

$$I_{d \text{ sat}} = (E_d - V_{res})/R_d \approx E_d/R_d \quad (8.45)$$

The operating point B of the on switch lies in the initial, quasilinear portion of the MOS transistor characteristic. Therefore, multiplying the saturation current of Eq. (8.45) by the channel resistance of Eq. (5.17) gives the residual voltage. Setting $V_{gs} = E_g^+$, we obtain

$$V_{res} = E_d/[b (E_g^+ - V_0) R_d] \quad (8.46)$$

where b is the specific transconductance given by Eq. (5.7).

Where switches operate jointly in a series network, the turn-on signal E_g^+ , comes from the preceding, off switch, in which case $E_g^+ = E_d$.

If we set $b = 0.1 \text{ mA/V}^2$, $V_0 = 2.5 \text{ V}$, $R_d = 50 \text{ k}\Omega$, and $E_g^+ = E_d = 7.5 \text{ V}$, then $V_{res} = 300 \text{ mV}$. This value of V_{res} is comparatively large and there are limited ways for its decrease in the given circuit because a growth both in R_d and in b leads to an increased area occupied by the circuit, which is undesirable in semiconductor circuits.

But it should be pointed out that there are no **principal** limitations on the quantity V_{res} in MOS transistor switches: *the residual voltage can be made as small as desired* by raising the resistance R_d and voltage E_g^+ . This is one of the most important advantages of MOS transistor switches over bipolar counterparts in which V_{res} is in principle limited by the voltage V_{ce} given by Eq. (8.2).

8.7.2. Dynamic-load switch. The circuit of such a switch using single-type transistors is shown in Fig. 8.17a. The dynamic load here is the transistor $T2$ with its gate connected to the drain, so that the transistor essentially plays the role of a resistor. In this switch network, the transistor $T2$ is called a *load* and $T1$ an *active* transistor.

The I - V characteristic of the resistor $T2$ can be plotted relying on the following considerations. Since we have $V_{gs2} = V_{ds2}$ when connecting the gate to the drain, the inequality $V_{gs2} - V_0 < V_{ds2}$ is obviously valid. This inequality, according to (5.5), means that the transistor $T2$ acts in the **flat** portion of the curve. For this portion,

Eq. (5.8) is true. Substituting $V_{gs} = V_{ds2}$ into this equation, the expression for the I - V characteristic takes the form

$$I_{d2} = 1/2 b_2 (V_{ds2} - V_{02})^2 \quad (8.47)$$

As clear, this I - V curve is parabolic, that is, nonlinear.

In the off switch, with the voltage $E_g^+ < V_0$ applied to the gate, the residual current is approximately the same as in a resistive-load switch (10^{-9} to 10^{-10} A or below), and the maximum output voltage is close in value to the supply voltage: $V_{\max} \approx E_d$ (see point A in

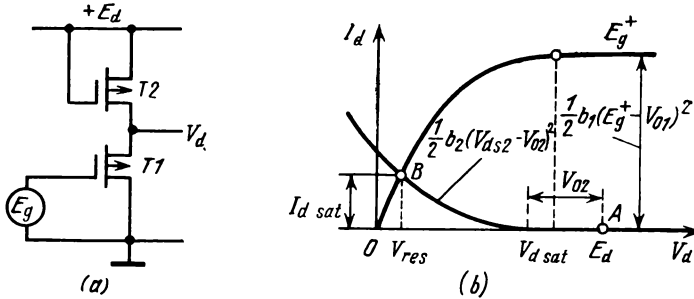


Fig. 8.17. MOS transistor switch with dynamic load
(a) circuit; (b) operating points on output characteristic

Fig. 8.17b). The exact position of point A is at the intersection of the reverse characteristics for the drain pn junctions of the active and load transistors.

At switch-on, when the voltage on the gate is $E_g^+ > V_0$, the quiescent point B lies in the quasilinear portion of the characteristic for the active transistor $T1$. The residual voltage at this point is small as usual. The supply voltage can thus be considered to be fully applied to the load transistor $T2$. The saturation current of the switch is defined by Eq. (8.47) if V_{ds2} is set equal to E_d :

$$I_{d \text{ sat}} = 1/2 b_2 (E_d - V_{02})^2 \quad (8.48)$$

Multiplying the current $I_{d \text{ sat}}$ by the channel resistance of Eq. (5.17) and assuming $V_{gs} = E_g^+$, we find the residual voltage in the form¹

$$V_{res} = \frac{b_2 (E_d - V_{02})^2}{2b_1 E_g^+ - V_{01}} \quad (8.49)$$

Because the condition $E_g^+ \leq E_d$ is always met in practice, the following important conclusion can easily be made: in order that the resi-

¹ The threshold voltages are taken different for generality. In integrated circuits this difference is inevitable due to the difference in voltages between the sources and the common substrate.

dual voltage might be small, *the condition $b_2 \ll b_1$ must be fulfilled in the dynamic-load switch*; in other words, *the transistors must be substantially different*.

Let us recall that the specific transconductance b is primarily the function of transistor geometry, namely, the channel width-to-length ratio Z/L , as is clear from Eq. (5.7). Hence, the ratio Z/L must be the largest possible in an active transistor, and the lowest possible in a load transistor. In both cases the limitations stem from design and manufacturing factors. If we achieve the ratio $b_1/b_2 = 50$ to 100, which is quite practicable, then the residual voltage can lie in the range from 50 to 100 mV.

8.7.3. Complementary switch. The circuit of this switch is illustrated in Fig. 8.18. Let in the initial state the control voltage be $E_g = 0$. Then

$$V_{gs1} = 0, \quad V_{gs2} = -E_d$$

Hence, assuming $E_d > |V_{02}|$, the n -channel transistor $T1$ is off and the p -channel transistor $T2$ is on.

The current in the common circuit is determined by the off transistor $T1$, and is equal to I_{res1} . As in the above switches, the on transistor $T2$ operates in the quasilinear region of the characteristic, where the channel resistance is described by Eq. (5.17). Multiplying the residual current of the first transistor by the channel resistance of the second and setting $V_{gs} = E_d$, we find the voltage in the on transistor $T2$:

$$|V_{ds2}| = \frac{I_{res1}}{b_2 (E_d - |V_{02}|)} \quad (8.50)$$

If we set $I_{res1} = 10^{-9}$ A, $b_2 = 0.1$ mA/V², and $E_d - |V_{02}| = 5$ V, then $|V_{ds2}| = 2\mu$ V.

Let now the control voltage go positive and take on a value $E_g^+ = E_d$, in which case

$$V_{gs1} = E_d > V_{01}, \quad V_{gs2} = 0$$

Now the n -channel transistor $T1$ is biased to conduct and the p -channel transistor $T2$ is driven into cutoff. So the current in the common circuit remains at a level of I_{res} , though the transistors have exchanged the operating conditions.

The most important feature of CMOS switches, as follows from the above description, is that they *do not practically consume power in both steady states*. These two states can thus be called "off" and "on"

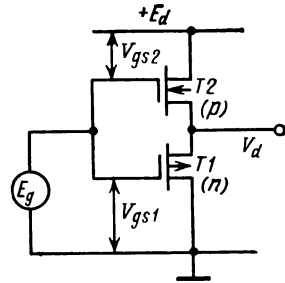


Fig. 8.18. Complementary MOS transistor switch

only conditionally with respect to **one** of the transistors, say, the n -channel transistor.

But both steady states differ rather sharply in the level of output voltage; as was shown above, at a low level of E_g^+ , when $T1$ is off, the voltage V_{ds2} is negligible, and hence the output voltage is equal to the supply voltage:

$$V_{\max} = E_d \quad (8.51a)$$

At a high level of E_g^+ , when $T1$ is on, the same negligible value of voltage drops across this transistor. The quantity V_{ds1} can be found from Eq. (8.50), replacing the indexes in the right side of the expression. This is exactly the residual voltage across the switch:

$$V_{res1} = I_{res2}/[b_1 (E_d - V_{01})] \quad (8.51b)$$

The residual voltage can take the same extremely small values as given in the above example, a few units of a microvolt and below. Low residual voltages are the second important advantage of complementary switches.

If the supply voltage E_d exceeds the sum of threshold voltages for both transistors, then we have an interval of control signals

$$V_{01} < E_g < E_d - |V_{02}|$$

within which **both** transistors are in the on condition. The current in the circuit will thus have a **finite** (and sometimes rather large) value, which can be calculated with the known formulas. CMOS switches, however, feature low supply voltages close in value to the sum of threshold voltages, so that a noticeable rise in current during switching does not generally occur.

8.7.4. Transients. The transient of MOS transistor switches is largely determined by the recharge of capacitances which form a part of the complex load. The channel response defined by the time constant τ_s of Eq. (5.27) can be allowed for, if necessary, by adding τ_s to the time constant of capacitance recharge.

Figure 8.19a illustrates the resistive-load switch using a transistor $T1$. The switch operates in a series circuit to drive a second similar switch. Fig. 8.19b gives the switch equivalent circuit which represents all individual capacitances by one total capacitance C_d :

$$C_d = C_g + C_{d\ sub} + C_{par} + C_{gs} + KC_{gd} \quad (8.52)$$

The typical values of the total capacitance C_d lie between 1 and 3 pF. It comprises the following components: the gate-to-channel capacitance C_g which, in distinction to other capacitances, is *inherent in the MOS transistor* as regards its principle of action [see Eq. (5.7) where the gate-to-channel per-unit area capacitance determines the value of specific transconductance]; the drain-substrate capacitance

$C_{d\ sub}$ (the barrier capacitance of the drain pn junction); the parasitic capacitance of wiring relative to the substrate, C_{par} (the capacitance of metallization in integrated circuits); and the capacitances C_{gs} and C_{gd} resulting from metal gate overlap (see Figs. 5.11 and 7.27). The factor K originates from the *Miller effect* (see Sec. 9.5) and can range from a few units to 10 to 20 and above, in which case the role of C_{gd} often becomes predominant.

Assume the transistor is initially on and a small residual voltage drops across it. The incoming cutoff signal E_g^- causes the current

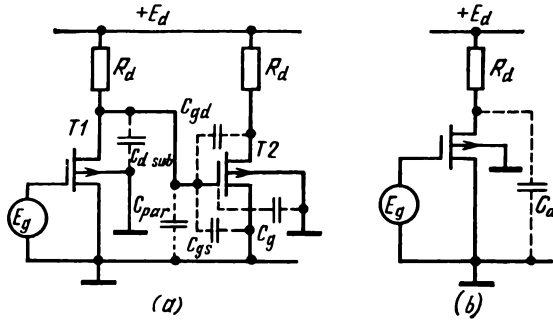


Fig. 8.19. Parasitic capacitances in a MOS transistor switch
(a) capacitance components; (b) resultant capacitance

in the transistor to decay to zero at a rate determined by a rather small time constant τ_s (practically instantaneously). After driving the transistor into cutoff, that is, after disconnecting the switch, C_d charges from the supply source E_d via R_d with a time constant $\tau_c = R_d C_d$ (Fig. 8.20a).

The process of charging obeys the simplest exponential function

$$v_d(t) = E_d (1 - e^{-t/\tau_c})$$

The charging time, or the rise time for the voltage at 90% level of E_d is

$$t_r = 2.3 R_d C_d \quad (8.53a)$$

Assuming $R_d = 50 \text{ k}\Omega$ and $C_d = 3 \text{ pF}$, we get $\tau_c = 150 \text{ ns}$ and $t_r \approx 350 \text{ ns}$.

If we replace R_d in Eq. (8.53a) by $E_d/I_{d\ sat}$ according to Eq. (8.45), the expression for rise time will assume a more general form

$$t_r = 2.3 (E_d C_d / I_{d\ sat}) \quad (8.53b)$$

It is clear that the rise time is primarily the function of the desired value of operating current.

The process of turning the switch off and forming the voltage-pulse trailing edge proceeds in a more complex way. After applying

a turn-on signal E_g^+ , the current I_d reaches the value

$$I_d(0) = 1/2 b (E_g^+ - V_0)^2 \quad (8.54)$$

with a time constant τ_s (Fig. 8.20b), or practically instantaneously. This current causes the capacitance C_d to discharge. In the process of discharging, the drain voltage V_d drops off. So long as this voltage keeps larger than the saturation voltage $V_{d\text{ sat}}$ equal to $E_g^+ - V_0$, the transistor operates in the flat region of the characteristic, and the current remains at a level of $I_d(0)$, as shown in Fig. 8.20c.

As V_d drops below the saturation voltage, I_d begins to decay, approaching $I_d(\infty) = I_{d\text{ sat}}$. At this stage of the transient, it would

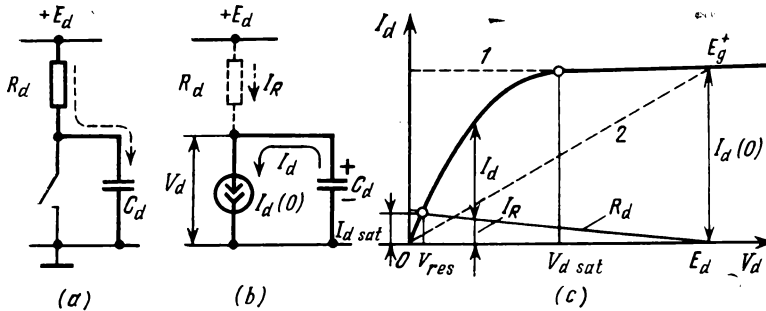


Fig. 8.20. An example of calculation of transients in a MOS transistor switch (a) equivalent circuit for switch in cutoff; (b) equivalent circuit for switch in on condition; (c) output characteristics showing the process of biasing into conduction

be necessary to allow for the nonlinear dependence $I_d(V_d)$. But since this leads to mathematical difficulties, we shall resort to two simplest approximations, one of which understates and the other overstates the fall time.

In both cases we shall disregard I_R through the load resistor, because over a greater length of the transient this current is small in comparison with I_C (see Fig. 8.20c).

As a first approximation, set $I_d = I_d(0) = \text{constant}$. The capacitance then discharges at an invariable current as shown by a dash line 1 in Fig. 8.20c. Dividing the initial charge $Q = E_d C_d$ by the discharge current $I_d(0)$ gives

$$t_f = E_d C_d / I_d(0)$$

As a second approximation, set $I_d = V_d / R_{av}$, where $R_{av} = E_d / I_d(0)$ is the average resistance during discharging (see dash line 2 in

Fig. 8.20c). In this case, discharging proceeds in a usual exponential manner:

$$v_d(t) = E_d e^{-t/\tau}$$

where $\tau = C_d R_{av}$. Assuming the discharge comes to an end at a 10% level of E_d yields

$$t_f = 2.3\tau = 2.3 [E_d C_d / I_d(0)]$$

For calculations, we can accept an intermediate of the above two values:

$$t_f = 1.5 [E_d C_d / I_d(0)] \quad (8.55)$$

where the current $I_d(0)$ is defined by Eq. (8.54). Thus, assuming $E_g^+ = E_d = 7.5$ V, $V_0 = 2.5$ V, $b = 0.1$ mA/V², and $C_d = 3$ pF, we find that $I_d(0) = 1.25$ mA and $t_f \approx 25$ ns.

As seen, the positive-pulse trailing edge is significantly shorter than the leading edge. In general, this conclusion follows from the structure of Eqs. (8.53b) and (8.55), which mainly differ in the values of current. From Fig. 8.20c it is evident that $I_d(0) \gg I_{dsat}$, and hence t_f is inevitably much smaller than t_r .

The switching speed of the given type of circuits is thus defined by the length of the leading edge. To shorten the time t_r , we need to decrease R_d , but this tends to raise the residual voltage on the switch as follows from Eq. (8.46). Consequently, there are limited possibilities for increasing the speed of response.

The general form of transients in the discussed switch circuit appears in Fig. 8.21.

In the dynamic-load switch shown in Fig. 8.17a, the trailing edge forms in the same way as in resistive-load switch, the fall time t_f being given by Eq. (8.55). This coincidence stems from the fact that in deriving (8.55) we have disregarded the load current I_R and hence the specifics of loading. The current $I_d(0)$ entering into Eq. (8.55) is here the initial current of the active transistor. By analogy to Eq. (8.54), we write

$$I_d(0) = 1/2 b_1 (E_g^+ - V_{01}) \quad (8.56)$$

As for the leading edge of the pulse, this builds up during charging of C_d via a **nonlinear** dynamic load. Considering the parabolic character of the I - V curve described by Eq. (8.47), we can expect the

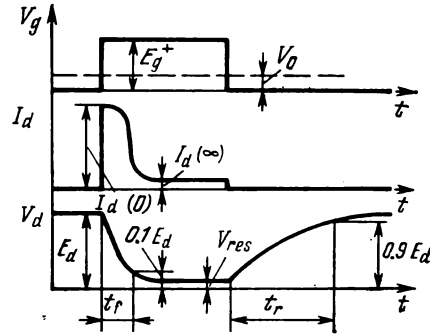


Fig. 8.21. Transients in a MOS transistor switch

capacitance to charge slower than would be the case with a resistive load, so that t_r will be greater. To obviate the need for complex mathematical calculations, we replace the parabolic I - V curve by the linear curve with a resistance $E_d/(1/2 I_{d\text{ sat}})$. Instead of Eq. (8.53b), we now have

$$t_r = 2.3 E_d C_d / (1/2 I_{d\text{ sat}}) \quad (8.57)$$

where $I_{d\text{ sat}}$ is the saturation current of Eq. (8.48).

Note that in the given switch the load capacitance of Eq. (8.52) should additionally contain C_{gs2} ; for integrated circuits, C_d should also include $C_{s\text{ sub}2}$ because the substrate is common to both transistors.

Since t_r has grown here in comparison with that in a resistive-load switch, while the time t_f has remained the same, we are in a position to conclude that in *switches* with a dynamic load, *the switching speed is defined by the rise time* as it is in switches with a resistive load.

Using Eqs. (8.47) and (8.56), it can be readily shown that the ratio t_r/t_f is primarily the function of b_1/b_2 . An attempt to reduce b_1/b_2 and thus to level off the leading and trailing edges of pulses leads to an increase in the residual voltage given by Eq. (8.49). For this reason a greater switching speed requires an increase in the specific transconductance of **both** transistors, but this, as known, entails an increase in their area. In integrated circuits, this approach naturally has its limits.

In a complementary switch, the specifics of transients lie in that the charging and discharging of load capacitance C_l occur *approximately under the same conditions* because of the symmetry of the circuit with respect to cutoff and turn-on input signals (see Subsec. 8.7.3).

The capacitance charges via the on transistor $T2$, with $T1$ switched off (see Fig. 8.18), and discharges via the on transistor $T1$ with $T2$ off. In both cases, the on transistor first operates in the flat portion of the I - V curves at a rather large current $I_d(0)$ and then, as the capacitance charges or discharges, the drain voltage drops below $V_{d\text{ sat}}$ and the current starts falling. So the mechanisms of charging and discharging are the same as those we have discussed in analyzing the discharge process in the switch with a resistive load (see Fig. 8.20b and c).

Consequently, the rise time and fall time depend on the same voltages as given in Eq. (8.55):

$$t_r = 1.5 \frac{E_d C_d}{I_{d2}(0)} = \frac{3E_d C_d}{b_2 (E_d - |V_{02}|)^2} \quad (8.58a)$$

$$t_f = 1.5 \frac{E_d C_d}{I_{d1}(0)} = \frac{3E_d C_d}{b_1 (E_d - V_{01})^2} \quad (8.58b)$$

The indexes 1 and 2 in Eqs. (8.58) have to stress the fact that the parameters of n - and p -channel transistors are different. But this difference is of secondary significance. The rise time and fall time are practically equal.

If, as done above, we set $V_0 = 2.5$ V, $b = 0.1$ mA/V², $C_d = 3$ pF, and $E_d = 7.5$ V, then $t_r = t_f \approx 25$ ns.

Comparing the found values with those given above, we see that the speed of a complementary switch is almost ten times that for the other two types. The same holds at a decreased supply voltage.

For all the three types of switch *the main way of raising the switching speed is to reduce the total capacitance C_d* . At a given capacitance, *the switching speed rises with currents*, in particular, with supply voltages.

8.8. Noise Immunity of Switches

Apart from valid (control) signals, spurious or unwanted signals always have an effect on the performance of switches. These parasitic or random signals result from external electromagnetic interferences (stray pickups) or from internal processes, such as coupling via a common supply source. The useful signals, therefore, must exceed the noise level, and a switch must be as insensitive to small parasitics as possible; in other words, the switch must not respond to undesired signals as readily as it does to useful signals.

A measure of insensitivity of a switch (or any other circuit) to noise is known as *noise immunity*.

It is customary to measure noise immunity in terms of the absolute value of a signal, commonly in volts, which does not yet cause false switching of the device to the on or off condition. The extent of immunity to positive and negative signals can differ substantially. The analysis of noise immunity presupposes that a switch operates in a series network.

Series-connected switches must function as a whole; a change in the state of the first switch must cause the changes in the states of the other switches, including the last switch. For this, the input signal must exceed the *sensitivity threshold*, otherwise the signals in the circuit would "decay" and the switches remote from the first would fail to change states. On the contrary, a noise signal must be below the sensitivity threshold.

To estimate the sensitivity threshold, consider first the general method for determining the operating points in the series network of switches. The statement of the problem is as follows: given the operating point of the n th switch, we have to find the operating point of the $(n + 1)$ th switch using the transfer characteristic. There are three approaches to solving this problem (Fig. 8.22). The consecutive

arguments rely on the obvious relation

$$V_{out\ n} = V_{in(n+1)} \quad (8.59)$$

where n is the switch number in the network.

Figure 8.22a displays the direct method that gives the answer to the problem. Assume we know the position of point n . Projecting this point onto the y -axis (arrow 1) gives the value of $V_{out\ n}$. Laying off this value on the x -axis (arrow 2), we obtain $V_{in(n+1)}$ according to Eq. (8.59). Last, drawing the vertical from this abscissa (arrow 3), we find the operating point $(n+1)$. We can repeat the procedure to find any other operating point.

The second method illustrated in Fig. 8.22b is more convenient and illustrative, and uses a bisector drawn from the origin of coordinates. The bisector is a geometric place of points characteristic of the equality $V_{out} = V_{in}$. So, projecting the point n on the bisector (arrow 1) gives the point n' whose abscissa is $V_{in(n+1)}$. Erecting now a perpendicular (arrow 2) from point n' until it intersects the transfer curve gives the point $(n+1)$. We shall use precisely this method in the further discussion.

The third method shown in Fig. 8.22c is similar to the load line approach (see footnote on p. 267). The transfer characteristic of the next switch, 2, is here a mirror image of the original curve, swung 90° to the curve of the preceding switch, 1. The point n on the curve 1 is projected along the horizontal onto the curve 2 (arrow 1) to give the operating point $(n+1)$. Next the point $(n+1)$ is projected along the vertical onto the curve

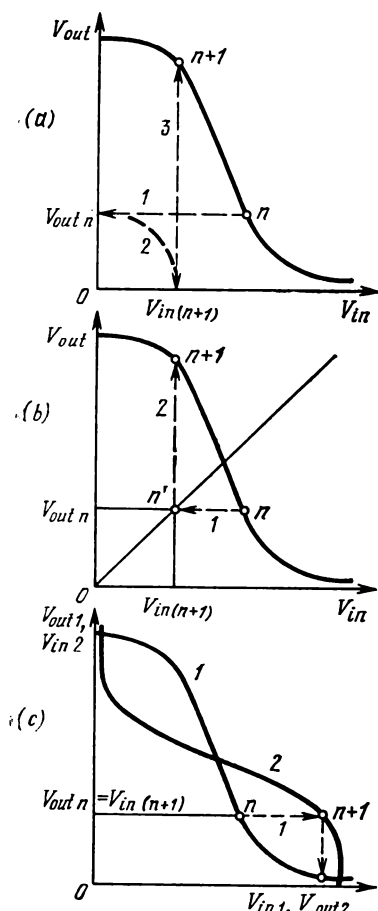


Fig. 8.22. Methods of determining operating points for the series network of switches

(a) direct method; (b) bisector method; (c) load-line method

1 (arrow 2) to give the point $(n+2)$. The procedure is then repeated to find other points.

Suppose that in the series circuit of Fig. 8.6 the switch $T1$ is off, that is, the voltage at its output is close to the supply voltage (point A in Fig. 8.1). The switch $T2$ is then on and its output has a residual voltage approaching zero (point B in Fig. 8.1). Accordingly, the switch $T3$ is off, $T4$ is on, and so forth. The output voltages of switches in the initial state are shown in Fig. 8.23a.

Apply now a signal V_{in1} to the input of switch $T1$ (Fig. 8.24a). The operating point of the switch will shift from the position A into a position 1 . Project the point 1 on the bisector and draw the vertical from point $1'$ as far as the transfer curve to obtain the operating point 2 for the switch $T2$. Project now the point 2 on the bisector and draw the vertical from the $2'$ until it intersects the transfer curve to give the operating point 3 for $T3$. Performing the procedure as before, we see that the operating points of odd switches, starting from $T5$, coincide with the point B (the on state), and the operating points of even switches, starting from $T4$, coincide with the point A (the off state). In other words, the pulse V_{in1} is sufficient to control the circuit because this pulse causes all the switches (excepting a few first switches) to change states (see Fig. 8.23b).

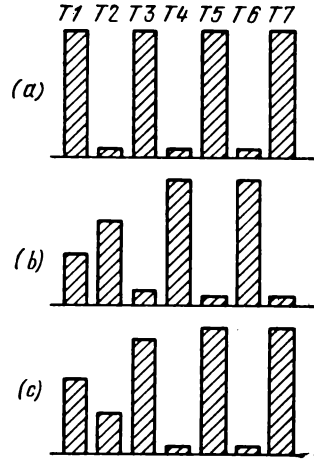


Fig. 8.23. Output voltages of switches connected in series

(a) initial condition; (b) when input signal is above threshold level; (c) when input signal is below threshold level

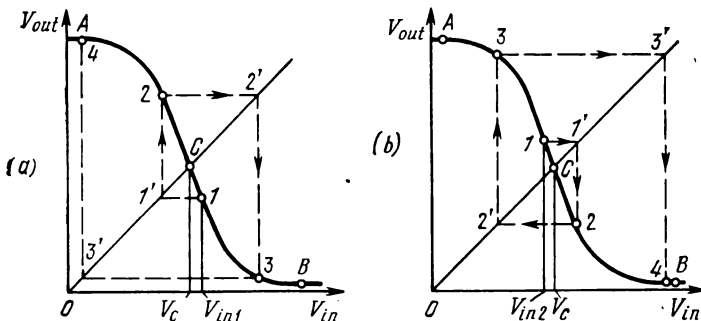


Fig. 8.24. Calculating the state of a switch in the circuit when applying a trigger input above (a) and below (b) sensitivity threshold

Consider now the case where a smaller control signal V_{in2} arrives at the input of the circuit being in the same initial state (Fig. 8.24b).

Carrying out the same procedure as described above, we obtain the operating points 1, 2, 3, and others, and find that the given signal is not large enough to effect control of the circuit because it changes the states of a few first switches (and merely partially), while the states of the rest of the switches remain the same (see Fig. 8.23c).

It is easy to see that the criterion of the sufficient signal value is $V_{in} > V_C$. Therefore, the voltage V_C corresponding to the point of intersection of the bisector with the transfer characteristic is called the *sensitivity threshold*.

At first glance, the sensitivity threshold directly determines the degree of noise immunity of the switching circuit; namely, if a positive noise pulse meets the condition $V_{in}^+ < V_C$, and a negative noise pulse the condition $V_{in}^- < (E - V_C)$,

the circuit state will not change, except for the states of a few first switches (Fig. 8.25)¹. In reality the above conditions are necessary but not sufficient.

The conditions sufficient to afford noise immunity (for reasons explained below) are:

$$V_{in}^+ < V_a, \quad V_{in}^- < E - V_b$$

Here V_a and V_b are the abscissas for points a and b for which the modulus of the derivative dV_{out}/dV_{in} is equal to unity. This derivative is nothing else than the *differential voltage gain* of a switch. If $V_{in}^+ < V_a$, the voltage gain $K < 1$, that is, the input signal does

not increase as it passes through the network, but decreases, and so there is no danger of false switching. But if $V_{in}^+ > V_a$, then $K > 1$ and the "top" of the pulse $V_{in}^+ - V_a$ is subject to amplification. **Where feedback is present**, this condition may become the cause of false switching even if the input signal V_{in}^+ is smaller than the sensitivity threshold. The same is true for negative noise signals.

Proceeding from the above reasoning, the extent of noise immunity of a switch can be estimated by the quantities

$$V_n^+ = V_a \quad (8.60a)$$

$$V_n^- = -(E - V_b) \quad (8.60b)$$

Both of these quantities are shown in Fig. 8.25.

¹ It is expedient to count off a negative noise pulse from the operating point of the **on** switch (point B in Fig. 8.1) since this pulse aids in biasing the formerly on switch into cutoff. This means that we have to compare V_n^- with $E - V_C$ rather than with V_C .

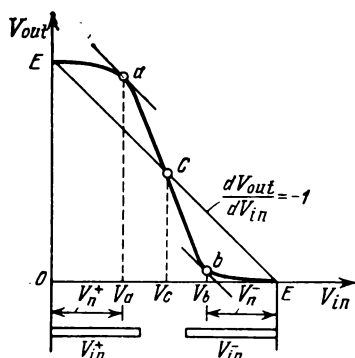


Fig. 8.25. Estimating the degree of noise immunity of a switch

8.9. Bistable Units and Flip-Flops

The simple switches discussed in the preceding sections form the basis of the entire digital circuit engineering. They find extensive uses as independent units (current choppers and various kinds of multiplexers) and also as components of various functional blocks, first of all, with binaries (bistable circuits). What distinguishes bistable units is that they use not only **direct** coupling between switches (as in a series network) but also positive **feedback** paths.

8.9.1. Circuit and the principle of action. In the series network of switches shown in Fig. 8.6, every switch has its neighbors biased to the opposite state. So, in any pair of the adjacent switches (T_n

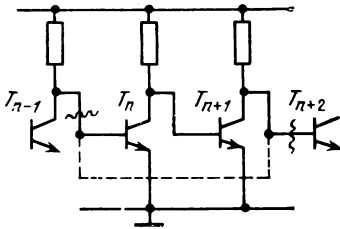


Fig. 8.26. Elements forming a bistable unit

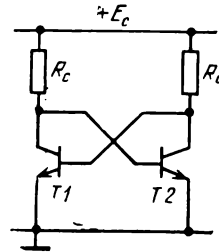


Fig. 8.27. A bistable unit

and T_{n+1} in Fig. 8.26), the output voltage of T_{n+1} is the same as the input voltage of T_n . Therefore, if we isolate these two switches from the preceding and consecutive switches and connect the output of the $(n+1)$ th switch to the input of the n th switch (see the dash line in Fig. 8.26), the pair in question will not change state. This stable state can be of two variants: T_n is on and T_{n+1} is off or, on the contrary: T_n is off and T_{n+1} is on. This type of electronic circuit having two stable states is called a *bistable unit* or *flip-flop*¹.

If we digress for a while from the "origin" of a bistable unit and represent it as an independent circuit (Fig. 8.27), we can see that this circuit features symmetric structure and crossed feedback paths. What characterizes the stable states of a bistable unit is that one of its switches is off and the other is on and biased into saturation. In other words, a bistable unit is noted for **electric asymmetry**. Let us show that electric symmetry for a bistable unit is impossible.

We shall construct the proof by contradiction. Assume the bistable circuit is in the symmetric state, so that both transistors (see

¹ Strictly speaking, the terms bistable unit and flip-flop are not synonymous. A bistable unit only forms the basis of any flip-flop which differs in the methods of control of a bistable unit (see below).

Fig. 8.27) are on and operated at the edge of the active region¹. The voltages on both collectors and both bases are equal and approach V^* ; the collector current is proportional to the base current

$$I_c = B I_b$$

Assume now that as a result of inevitable fluctuations (either internal or external), the voltage on one of the bases, for example, on the base of $T1$ has changed by a small value ΔV_{b1} . The currents will then change in the following manner:

$$\Delta I_{b1} = \Delta V_{b1}/R_{in}, \quad \Delta I_{c1} = B \Delta I_{b1}$$

Here R_{in} is the input resistance of the on transistor. A fraction of increment ΔI_{c1} will branch off to go into the base circuit of $T2$. Then,

$$\Delta I_{b2} = -m \Delta I_{c1}, \quad \Delta I_{c2} = B \Delta I_{b2}$$

where $m < 1$. In a similar way, a fraction of increment ΔI_{c2} will branch off into the base circuit of $T1$ to become an **additional** increment of its base current:

$$\Delta I'_{b1} = -m \Delta I_{c2} = m^2 B^2 \Delta I_{b1}$$

At typical values of m of about 0.5, the added increment $\Delta I'_{b1}$ that has appeared in passing around the circuit will be much higher than the initial increment ΔI_{b1} . The next increment $\Delta I''_{b1}$ will become as many times high as $\Delta I'_{b1}$, and so on. Hence, *the response of the circuit to the smallest initial fluctuation will result in its amplification*.

The avalanche-like current rise in one half of the bistable circuit and the corresponding current reduction in its other half is known as *regeneration*. The process of regeneration comes to an end after driving one of the switches into cutoff and the other into saturation. In the example under discussion, a **positive** fluctuation ΔV_{b1} switches the transistor $T2$ off, and a **negative** fluctuation renders $T1$ conductive.

Since the polarity of fluctuation is a random value, the results of the avalanche process (switch-off of $T2$ or $T1$) are equiprobable. So, in the analysis of a bistable circuit, we can regard any of the two stable states as the initial state.

The aim of triggering a flip-flop is to **set** the circuit in either of the two stable states by applying appropriate external signals, or **reset** the circuit from the given stable state to the opposite state. There are two methods of triggering a flip-flop circuit: *asymmetrical*

¹ The off condition for both transistors is impossible. Should this be the case, the collector potentials would be equal to $+E_c$ and thus would exceed the voltage V^* . The saturated condition for both transistors is likewise impossible, otherwise the collector potentials would be lower than V^* .

(set-reset) *triggering* that uses two trigger inputs and *symmetrical* (complementing, or count) *triggering* that uses only one (common) trigger input.

8.9.2. Asymmetrical triggering. In the asymmetrically triggered flip-flop of Fig. 8.28, one more transistor switch ($T3$ or $T4$) is connected in parallel with each of the transistors forming the bistable unit. These switches are under control of external signals, i.e., the base currents which assume one of the two values, I_b^+ or 0. Control switches perform the same functions as metallic contacts; they can be on or off.

Assume the initial state of the flip-flop is such that the transistor $T1$ is off, $T2$ is biased on to saturation, and both switches $T3$ and $T4$

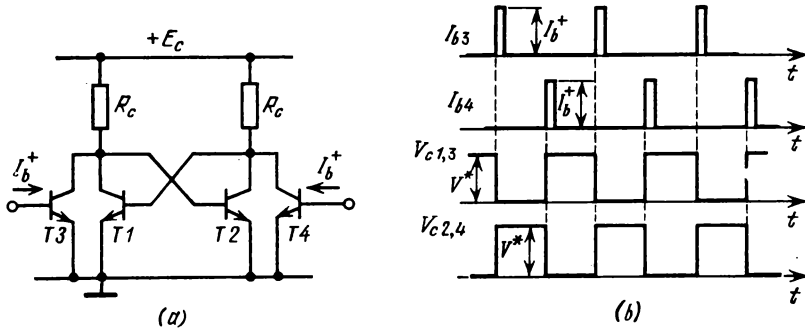


Fig. 8.28. Circuit (a) and time diagrams (b) of asymmetrically triggered flip-flop

stay off. If the positive-going step input I_b^+ drives $T4$ on to saturation, the state of the circuit will not change because V_{c2} in the initial condition has been close to zero. If now a step input causes $T3$ to conduct, the potential V_{c1} drops to zero and so does the base potential V_{b2} , with the result that $T2$ switches off. Then the process of regeneration pushes the transistor $T1$ into saturation. In its new stable state, the switch $T3$ does not any longer exert a control action on the circuit; its on and off states do not alter V_{c1} and V_{b2} . To return the flip-flop to the initial state, a trigger pulse must be applied to $T4$.

In asymmetrical triggering, therefore, trigger pulses alternately go to both inputs of the flip-flop (Fig. 8.28b). It should be pointed out that *the simultaneous arrival of trigger pulses at both inputs of the asymmetrically triggered flip-flop is impermissible.*

To illustrate the point, assume that both trigger pulses act simultaneously. In this case, the bases of $T1$ and $T2$ will be at a zero potential, and so both transistors will be off. After cessation of the pulses,

both transistors will turn on, and so the bistable circuit will **temporarily** stay in the symmetric condition. As shown above, the circuit can switch over to either of the two stable states with equal probability. So *the result of the simultaneous action of trigger pulses proves ambiguous*, which is unacceptable in digital circuits. The discussed type of bistable circuit with asymmetrical triggering received the name of a reset-set (RS) flip-flop. A *set* input signal causes the circuit to become set and a *reset* input causes it to reset.

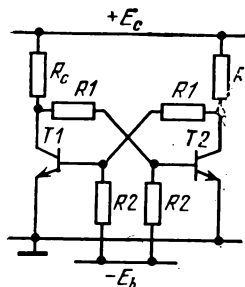


Fig. 8.29. Symmetrically triggered flip-flop with a cutoff voltage source

The structure of RS flip-flop circuits based on germanium transistors which were in use in the 50s and at the start of 60s was more complex than the silicon transistor circuit described above. This is because germanium transistors show the same value of residual voltage in the saturation region as silicon transistors (this voltage is independent of I_{c0}), but have a much smaller value of V^* , typically 0.2 or 0.3 V. So it is impossible to drive one of the transistors into cutoff by the residual voltage of the second on-transistor. To avoid this difficulty, feedback circuits had to incorporate

voltage dividers $R1$, $R2$ and a cutoff emf source $-E_b$ (Fig. 8.29).

The emf $-E_b$ ensures a negative potential on the base of the off transistor. In an electron tube version, such a flip-flop known as a "cathode relay" was suggested by the known Soviet radio engineer M. A. Bonch-Bruevich as far back as 1919.

8.9.3. Symmetrical triggering. In this type of triggering, each consecutive pulse applied **simultaneously** to the two interconnected inputs (forming a common input) causes the flip-flop to change its state to the opposite.

We have noted earlier that in the simple circuit of Fig. 8.28a, it is impermissible to apply trigger pulses to the two inputs simultaneously because after pulse cessation the circuit will change state irregularly. In order that the states of the bistable circuit might change **regularly** after each incoming pulse, the circuit must have an *internal memory*. The function of this memory is to store information on the preceding state of the circuit during the trigger pulse action and to force the circuit to change its state to the opposite after pulse cessation.

A classical approach to ensuring an internal memory is to employ commutating or *memory capacitors*. A flip-flop employing a capacitive memory and operating waveforms appears in Fig. 8.30.

Assume that in the initial condition the transistor $T1$ is biased off and $T2$ biased on to saturation. Hence,

$$I_{b1} = 0, \quad I_{b2} = (E_c - V^*)/(R + R_c)$$

The voltages on capacitors $C1$ and $C2$ will then be given by

$$V_{C1} = I_{b2}R = (E_c - V^*) \frac{R}{R + R_c}$$

$$V_{C2} = I_{b1}R = 0$$

and the collector potential of $T1$

$$V_{c1} = V^* + I_{b2}R$$

A trigger pulse applied to $T3$ and $T4$ causes both transistors to switch on to saturation. The potential V_{c1} then drops to practically zero, while the potential V_{b2} goes negative:

$$V_{b2} = V_{c1} - V_{C1} \approx -V_{C1}$$

Transistors $T1$ and $T2$ both stay off until the end of the input pulse. The incoming pulse causes the capacitor $C1$ to discharge via the re-

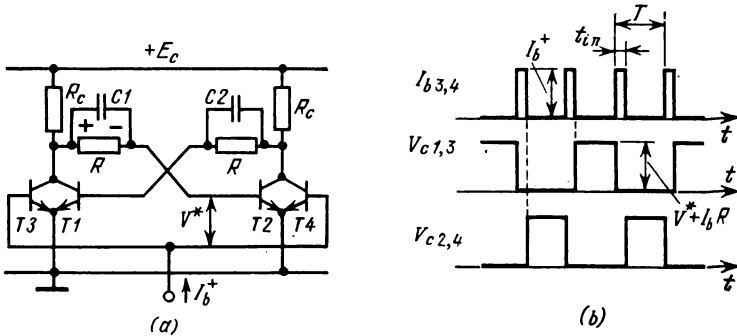


Fig. 8.30. Circuit (a) and typical waveforms (b) for a symmetrically triggered flip-flop

sistor R with a time constant $\tau_c = C_1R$. If the input pulse is rather short ($t_{in} \ll \tau_c$) the capacitor discharge is negligible and the voltage V_{C1} remains at its initial level.

As the input pulse ceases, turn-on currents which substantially differ in value start flowing into the bases of transistors:

$$I_{b1}^+ = \frac{E_c - V_{C2} - V^*}{R_c} = \frac{E_c - V^*}{R_c}$$

$$I_{b2}^+ = \frac{E_c - V_{C1} - V^*}{R_c} = \frac{E_c - V^*}{R_c + R}$$

It is obvious that I_{b1}^+ exceeds I_{b2}^+ . The initial rate of rise in the collector current of $T1$ will thus be higher than this is the case for $T2$.

The rapidly growing current I_{c1} branches off into the base of $T2$ and causes the initial value of I_{b2}^+ to go to zero. The transistor $T2$ then becomes off, while the transistor $T1$ enters the saturation region in a certain time. **The flip-flop has thus changed its initial state.**

In the interval between the input pulses the capacitor $C1$ has time to discharge and the capacitor $C2$ to charge to the same voltage as that which was on $C1$ in the initial condition. The next input pulse will trigger the processes similar to those described above, and so the flip-flop returns to its original state.

The capacitors $C1$ and $C2$ retain (remember) the voltages specific to the preceding state; they thus ensure *unambiguous artificial asymmetry of turn-on currents* when the input pulse ceases, and hence enable the circuit changeover from one state to another. A bistable unit whose action is to change state every time an input pulse is applied to the single (common) input is called a *toggle (T-type) flip-flop*, or *complementing flip-flop*.

For the normal operation of the considered T flip-flop, two conditions need to be met. One of the conditions mentioned earlier states: $t_{in} \ll \tau_c$, where τ_c is the time constant of capacitance. This condition enables retaining the charge on the commutating capacitor during the action of the input pulse. The second condition enables the capacitor discharge in the interval between input pulses. This condition reads: $T > 3\tau_c$, where T is the pulse repetition rate (the time measured from the leading edge of one pulse to that of the next pulse as shown in Fig. 8.30b). The above condition limits the speed of a switching circuit.

Since capacitors are undesirable elements in semiconductor ICs, integrated T flip-flops do not use memory capacitors. The required internal storage is effected by other specific means (see Ch. 10) which, besides, ensure a high flexibility of the circuit and additional functional possibilities.

8.9.4. Transients. It stands to reason that the circuit changeover from one stable state to the other does not occur instantaneously, since every switch entering into the circuit shows a response lag with changes of currents and voltages.

In asymmetrical triggering, the transient proceeds in the following way. Let in the original state the transistor $T1$ be off and $T2$ on (see Fig. 8.28). Assume at a moment $t = 0$ the incoming step input abruptly drives the control switch $T3$ on to saturation. At the first stage of the transient the transistor $T2$ begins to switch off as a result of excess charge dissipation in its layers and then the collector current rapidly drops to zero (see Subsec. 8.4.5). At the second stage, the input capacitance of $T1$ starts charging. When the voltage V_{b1} reaches V^* , the transistor $T1$ becomes on. At the third stage, the current I_{c1} starts to grow, and the stage culminates in saturation of the

transistor $T1$. At the fourth stage, the excess charge builds up in the layers of the saturated transistor (see Subsec. 8.4.3).

From the above we can conclude that the total switching time t_{sw} for an asymmetrically triggered flip-flop comprises such basic time components as storage time t_s , delay time t_d , rise time t_r , and charge accumulation time t_{ac} :

$$t_{sw} = t_s + t_d + t_r + t_{ac} \quad (8.61)$$

As regards the expressions for each of the time components, see Section 8.4.

Let us recall that if the input pulses are rather short, the charge accumulation time according to Eq. (8.22) depends on the duration of an input pulse. In the limit, if $t^+ = t_d + t_r$, the excess charge accumulation does not occur, and so $t_{ac} = 0$. So the maximum operating frequency of a flip-flop is defined by the switching time if $t_{ac} = 0$:

$$F_{\max} = (t_s + t_d + t_r)^{-1} \quad (8.62)$$

Setting $t_s = 5$ ns, $t_d = 3$ ns, and $t_r = 2$ ns, we have $F_{\max} = 100$ MHz.

In symmetrical triggering, the transient proceeds in a somewhat different manner. For the entire length of a trigger pulse, both transistors stay in the off condition. After cessation of the input pulse one of the transistors which was off in the original state begins to switch on. This transistor stores an excess charge. Concurrent with the excess charge buildup, another process continues, which involves **recharge** of the memory capacitors in compliance with the new steady state of the flip-flop (see p. 305).

The length of the first of the mentioned stages is obviously equal to t_{in} . The length of the second stage includes t_d and t_r . Last, the duration of the third stage is the time taken for the capacitors to recharge; this is the *recovery time* t_{rec} , which can be taken to be equal to $3\tau_c$, where $\tau_c = CR$.

The total switching time can be written thus:

$$t_{sw} = t_{in} + t_d + t_r + t_{rec} \quad (8.63)$$

As mentioned earlier, the length of an input pulse must satisfy the condition $t_{in} \ll \tau_c$. On the other hand, the pulse length must not be smaller than the storage time for a saturated transistor, otherwise the transistor would remain in saturation and the flip-flop would not change state. Considering both limitations, we set $t_{in} \approx t_s$. Then the first three summands of Eq. (8.63) will approximately be the same in value as in Eq. (8.61) for asymmetrically triggering. Hence, we can write the maximum operating frequency for a symmetrically triggered flip-flop in the form

$$F_{\max} = (t_s + t_d + t_r + t_{rec})^{-1} \quad (8.64)$$

It is easy to see that *in asymmetrical triggering the minimum switching time is larger and the maximum frequency is smaller than in symmetrical triggering*. Thus if we take t_s , t_d , t_r to be the same as in the above example and set $C = 10$ pF, $R = 1$ k Ω , then $t_{rec} = 3\tau_c = 30$ ns and $F_{max} \approx 25$ MHz, or one-fourth the frequency for asymmetrical triggering.

8.10. Schmitt Trigger

The Schmitt trigger is essentially a bistable pulse generator, or current switch, discussed in Sec. 8.6. To stress this fact, the Schmitt trigger circuit of Fig. 8.31 shows the current switch by solid lines and the voltage divider $R1$, $R2$ by dash lines. For the same purpose, the potential V_{b2} is denoted as E , though E is not constant in the given case.

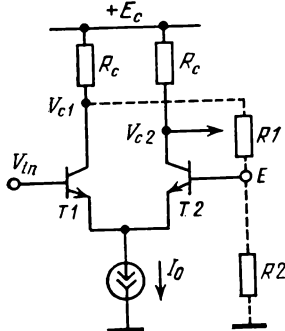


Fig. 8.31. Schmitt trigger

To simplify the analysis we shall idealize the $R1$ - $R2$ divider, assuming that it does not draw current and only conveys a fraction of voltage, V_{c1} , to the base of transistor $T2$:

$$E = \gamma V_{c1}, \quad \text{where} \quad \gamma = R_2 / (R_1 + R_2)$$

Suppose that in the original state the transistor $T1$ is off and $T2$ is on, but remains in the active region. For this state, the following potentials are valid:

$$V_{c1} = E_c, \quad E = \gamma E_c, \quad V_{c2} = E_c - I_0 R_c$$

With $T2$ in the active region, the condition $V_{c2} \geq E$ should be met, whence the limiting value of $I_0 R_{c2}$ can readily be determined.

The initial condition keeps invariable so long as $T1$ remains off, in which case input voltages remain smaller than E . Let us denote the voltage that drives $T1$ on by V_{in}^+ :

$$V_{in}^+ = E - \delta = \gamma E_c - \delta \quad (8.65)$$

where $\delta \approx 0.1$ V. If $E_c = 5$ V and $\gamma = 1/2$, then $V_{in}^+ \approx 2.4$ V.

Let the input signal V_{in} be slightly in excess of the turn-on voltage to cause a small current increment ΔI_{c1} to appear. This leads to the following train of events:

$$\Delta V_{c1} = -\Delta I_{c1} R_c, \quad \Delta E = \gamma \Delta V_{c1}, \quad \Delta V_e = \Delta E$$

The last equality presupposes that the forward voltage across the emitter junction does not vary and equals V^* . The increment ΔV ,

causes an **additional** increment in collector current:

$$\Delta I'_{c1} = -S \Delta V_e = \gamma S R_c \Delta I_{c1}$$

where S is the transconductance of the transistor¹.

If the product $\gamma S R_c$ is in excess of unity, the additional increment $\Delta I'_{c1}$ that arises in passing around the closed circuit will be larger than the initial increment ΔI_{c1} . This means that a *regenerative* (avalanche-like) process develops in the circuit, which causes the current I_0 to flow into the transistor $T1$ and the transistor $T2$ to switch off.

The voltage V_{in}^+ which causes an abrupt switchover of current I_0 from the transistor $T2$ to $T1$ is called the *upper trigger level* (UTL) for the trigger circuit.

After the switchover, the potentials in the circuit become

$$V_{c1} = V_{c0}, \quad E = \gamma V_{c0}, \quad V_{c2} = E_c$$

where V_{c0} is the collector potential of the transistor $T1$ biased on. Depending on the parameters of I_0 and R_{c1} the on transistor $T1$ can be operated both in the active and in the saturation region. The operation in the active region is typical. Then,

$$V_{c0} = E_c - I_0 R_{c1} \quad (8.66)$$

where $V_{c0} \geq V_{in}^+$ since the base potential keeps invariable in switching.

To return the circuit to the original state, we should reduce the input signal to a value close to E at which the transistor $T2$ starts to switch on.

Designate the turn-on voltage of $T2$ as V_{in}^- :

$$V_{in}^- = E + \delta = \gamma V_{c0} + \delta \quad (8.67)$$

If we set $\gamma = 1/2$ and $V_{c0} = V_{in}^+ = 2.4$ V according to Eq. (8.65), then $V_{in}^- \approx 1.3$ V.

As soon as the voltage V_{in} drops below V_{in}^- , the transistor $T2$ begins to go on and regeneration again occurs. The circuit then abruptly returns to the initial condition at which $T1$ is off and $T2$ is on.

¹ The minus sign in the relation between the increments ΔI_{c1} and ΔV_e is due to the fact that in *npn* transistors the current rises with a *negative* increment in emitter potential.

The level of input voltage which causes an abrupt switchover of current I_0 from $T1$ to $T2$ is known as the *lower trigger level* (LTL).

Disregarding the small value of δ and taking into account the equality $V_{c0} = V_{in}^+$, from Eq. (8.67) we find that *the LTL is smaller in magnitude than the UTL*. This relationship is of primary importance for the Schmitt trigger.

The terms *upper trigger level* and *lower trigger level* are often replaced by more general terms *upper threshold level* and *lower threshold level*. The Schmitt trigger is therefore called a threshold device.

The output signal is derived from the collector of $T2$. Since this collector is free of the feedback path and the transistor $T2$ is operated in the nonsaturated region, the transients take an extremely small time, much smaller than in ordinary flip-flops. This is one of the most important advantages of the Schmitt trigger over other bistable circuits.

Figure 8.32 shows the transfer characteristic for a Schmitt trigger, where $V_{out} = V_{c2}$. As seen, the difference between the UTL and LTL

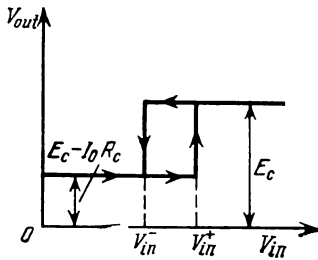


Fig. 8.32. Transfer characteristic for a Schmitt trigger circuit

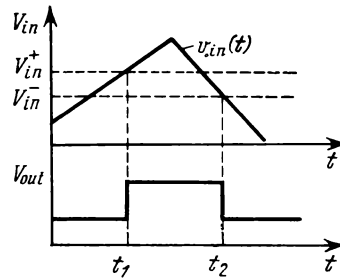


Fig. 8.33. Wave input and output of a Schmitt trigger circuit acting as voltage level detector and pulse shaper

results in the hysteresis of the circuit, the loop width being $V_{in}^+ - V_{in}^-$. For the above values of V_{in}^+ and V_{in}^- , the amount of hysteresis comes to about 1.1 V.

A typical method of using the Schmitt trigger as a threshold element is illustrated in Fig. 8.33. As V_{in} rises smoothly from zero, the trigger circuit remains in the initial state, with $T1$ off. As V_{in} reaches V_{in}^+ , the trigger changes state, the transistor $T2$ becomes off and a positive voltage step appears at the output. If the input signal does not reach an upper threshold level, the output signal does not appear. The Schmitt trigger can thus sort out input signals by their amplitude: above or below the threshold level V_{in}^+ . Such a circuit is called a *pulse-height analyzer* or *discriminator*.

If the amplitude of input signals is known to be higher than the threshold level, the function of the Schmitt trigger changes: the

circuit becomes a *pulse generator* which converts **smoothly varying** input signals to standard-height pulses with sharply shaped edges. Such pulse-shaping circuits are often necessary for various practical applications.

In conclusion it can be pointed out that rather high input signals ($V_{in} > V_{in}^+$) can drive the transistor *T1* into saturation. This does not trigger *T2* to the on state and this does not disturb the performance of the circuit, but causes a sharp rise in the input current of the trigger and slows down its switching speed because of the need for removal of the stored excess charge.

9.1. General

In Sec. 8.2 we have outlined the general features of analog circuits whose range is rather large and diverse. Until the advent of microelectronics one could hardly think that it was possible to find a few standard analog circuits like switches in digital engineering, that would serve as the basis for all or most of the analog circuits. As microelectronics developed further, however, the search for such standard circuits continued and proved a success. These circuits are briefly described in the next chapter.

Amplifiers always played a leading role in analog circuit engineering. They performed the most general function of amplifying the power of weak signals. In microelectronics, their role has become yet more important because the so-called operational amplifier (the basic variety of analog ICs) can serve most different purposes, of which amplification is only one of the many. That is why we shall primarily focus on amplifiers and consider them most thoroughly.

Amplifiers can be classified in a variety of ways. By the band (range) of frequencies being dealt with, the amplifiers are divided into *wide-band* and *narrow-band (selective)* types. The former include an important group of *dc amplifiers* capable of amplifying signals of the lowest possible frequency. By the power handled, the amplifiers can be grouped into low-power and high-power categories. In the first category, the power of an output signal is much lower than the power consumed by an amplifier; in the second, both powers are comparable in value. By the purpose they have to serve, the amplifiers can be classified as linear, logarithmic, differentiating, electro-metric, and others.

Discrete transistor circuit engineering widely used the classification of amplifiers by the types of coupling between individual stages, such as capacitive, transformer, and conductive (direct) coupling. Capacitive coupling was most popular and the study of amplifiers began with the treatment of this type. But, as known, the manufacture of capacitors in semiconductor ICs involves difficulties, while the manufacture of transformers is impossible. This explains why direct-coupled amplifiers, considered "exotic" in discrete circuit engineering, have come to be in the forefront.

Since direct (nonreactive) coupling enables the amplification of arbitrarily slow signals, the amplifiers with such coupling are dc amplifiers mentioned above. They form the basis of integrated analog engineering.

9.2. Composite Transistors

Many analog ICs use a few (typically, two) transistors connected so that they can be regarded as a single component called a *composite transistor*. Composite transistors exhibit the properties that are difficult or impossible to achieve in individual transistors of a conventional structure.

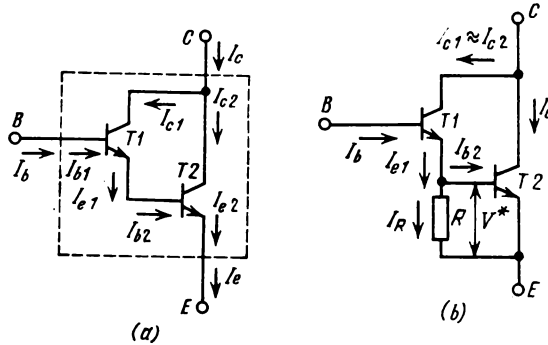


Fig. 9.1. Darlington pair
(a) simple; (b) current-equalizing

Among composite transistors, the most popular type is a *Darlington pair* (Fig. 9.1). The main feature of the Darlington pair is that it ensures an exceptionally high base current gain.

Indeed, from Fig. 9.1a it follows that

$$I_{b2} = I_{e1} = (B_1 + 1) I_b, \quad I_c = B_1 I_b + B_2 I_{b2}$$

Substituting the expression for I_{b2} into the second equality and dividing both sides of the equation by I_b gives the equivalent gain of the Darlington pair:

$$B = B_1 + B_2 + B_1 B_2 \quad (9.1a)$$

In all practical cases, the first two terms on the right of the expression are insignificant, and so the equivalent current gain can be written as

$$B = B_1 B_2 \quad (9.1b)$$

If the components B_1 and B_2 are equal to 100 to 200, the calculated gain B reaches 1×10^4 to 4×10^4 . The differential (dynamic) gain β will be approximately of the same value.

It should be borne in mind, however, that the transistors comprising the Darlington pair operate in fairly different modes: the emitter current I_{e2} exceeds about B_2 times the current I_{e1} . Considering the B - I_e relation shown in Fig. 4.11a, we conclude that B_1 can be much smaller than B_2 . The real values of B range into a few thousands, as they do in superbeta transistors discussed in Subsec. 7.4.4.

To balance out the currents I_{e1} and I_{e2} , a resistor R is connected in parallel with the emitter junction of transistor $T2$ (Fig. 9.1b). The current I_R through the resistor is equal to approximately V^*/R . This current can be made to approach the current I_{e1} , so that I_{b1} will be a small fraction of I_{e1} . In this case, the emitter and collector currents of both transistors will be nearly equal. Correspondingly, the static (dc) current gain B decreases to $2B_1$. But the ac current gain β can be very large as before. In comparison with B given by Eqs. (9.1), the gain β will decrease in the ratio of R/R_{in} , where R_{in} is the input resistance of $T2$. Calculations show that β decreases by

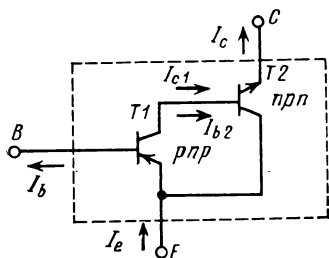


Fig. 9.2. Composite *pnp* transistor

a factor of merely 3 to 8, so its values lie in the range between 1 000 and 5 000. The parameters r_e and r_c of the Darlington pair are close in value to the respective parameters of transistor *T1*.

In Fig. 9.2 is illustrated the circuit of another composite transistor which can be called a *composite pnp transistor*. The given structure represents a combination of two transistors of the *pnp* and *nnp* types. As is clear from the figure, the resultant currents flow in the directions typical for the *pnp* transistor. As for the current gain of this pair, its expression $B = B_1 + B_1B_2$ practically resembles expressions (9.1) for the Darlington pair.

As noted in Sec. 7.5, integrated *pnp* transistors are inferior to *npn* transistors in current gain β . Thus the composite *pnp* transistor offers definite advantages because its current gain exceeds that of the *nnp* transistor entering into the common circuit of the pair. It is thus possible to use, say, a parasitic *pnp* transistor of Fig. 7.20a with its low current gain as a transistor $T1$ in the pair. The speed of response of the composite transistor is the same as for the constituent *pnp* transistor, that is, slower than for the *nnp* transistor.

9.3. Statics of a Simple Amplifier

Amplifiers generally consists of a few elementary cells known as *amplifying stages*. An amplifying stage may contain one, two, or more transistors, so its circuit may be rather complex. But even a

very complex stage cannot be divided into simpler components without losing its specific properties. In this sense, a single stage is an elementary cell of the amplifier.

9.3.1. Circuit and quiescent state. A simple direct current amplifier using one transistor is shown in Fig. 9.3a. This type of amplifier requires a dual \pm power supply (dual-polarity power supply), that is, two power sources of voltages $+E_c$ and $-E_e$ with respect to ground.

In principle, an amplifier can operate from single-polarity power supply (Fig. 9.3b), though this entails considerable difficulties. First, the circuit requires a special bias source E_b . Second, the

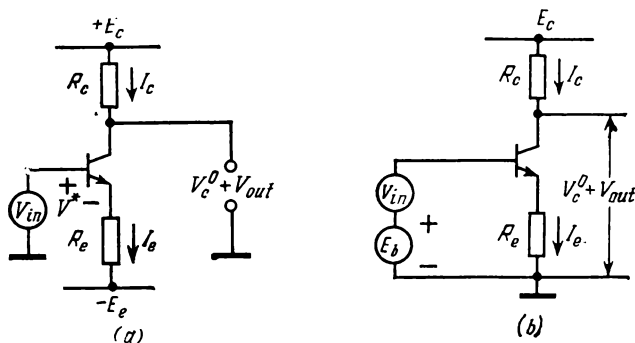


Fig. 9.3. Amplifying stages with dual-polarity (a) and single-polarity (b) power supply

signal source has no **grounded** point, which excludes the use of a number of typical signal sources and sharply raises the level of picked-up noise at the amplifier input. If we exchange the signal source and the bias source in the circuit, a high level of noise remains the same. Besides, it proves practically impossible to obtain a non-grounded bias source.

Let an input signal V_{in} be equal to zero. In this case, the power supplies E_c and E_e will cause *dc components* to flow in the circuit. The condition at which no input signals are present is commonly referred to as a *quiescent state*.

If an input signal is present, *ac components* proportional to V_{in} will add to the dc components. So, in the operating condition, the total voltages and currents can be written in the form: $V = V^0 + \Delta V$; $I = I^0 + \Delta I$. Here the upper index "0" identifies dc components and the increments stand for the ac components which are usually small as against the dc components.

As mentioned in Sec. 4.6, it is possible to analyze dc and ac components separately. Consider dc components typical of the quiescent state.

Set $V_{in} = 0$ and construct an amplifier circuit model such as shown in Fig. 9.4. The transistor equivalent circuit here is a simplified Ebers-Moll model corresponding to the normal active mode of operation (compare with Fig. 4.13), with a resistance R_b inserted into the base circuit for generality. This resistance comprises an internal base resistance r_b and also the resistance of the source resistance or that of the preceding stage.

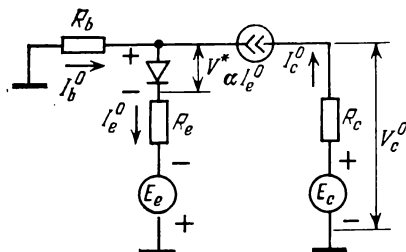


Fig. 9.4. Amplifying stage circuit model for dc components

Tracing around the input section of the circuit model of Fig. 9.4 gives

$$I_b^0 R_b + V^* + I_e^0 R_e - E_e = 0$$

Substituting $I_b^0 = (1 - \alpha) I_e^0$, we can readily determine the emitter current

$$I_e^0 = \frac{E_e - V^*}{R_e + (1 - \alpha) R_b} \quad (9.2a)$$

The collector potential has the form

$$V_c^0 = E_c - I_c^0 R_c \quad (9.2b)$$

where $I_c^0 = \alpha I_e^0$.

The quantities I_e^0 and V_c^0 have specified values. The combination of these values determines what is termed an *operating point* of the transistor in the quiescent state. The voltage E_c is commonly specified too; expression (9.2b) then unambiguously gives the required resistance R_c .

As for E_e and R_e , they both must be sufficiently large in order that inevitable changes in α and V^* might not have a noticeable effect on I_e^0 . It can be said that the choice of the values of E_e and R_e is determined by the **desired stability** of the operating point for the transistor with changes in temperature and other factors.

The choice of R_e is made proceeding from the condition

$$R_e > (1 - \alpha) R_b \quad (9.3)$$

Thus, if $R_b = 2 \text{ k}\Omega$ and $\alpha = 0.99$ (that is, $B = 100$), then R_e must be not less than 50Ω .

Having chosen R_e , we can easily find the quantity E_e from Eq. (9.2a). It may happen that the value of E_e is not large enough to prevent the effect of changes of V^* . Where this is the case, it is necessary to increase E_e and also R_e . Commonly, $E_e > 2$ or 3 V.

9.3.2. Differential parameters. A signal V_{in} causes changes in the voltages and currents in the circuit, that is, “gives” birth to ac components. To evaluate these components, let us use the small-signal transistor circuit model of Fig. 4.16. Restricting ourselves for

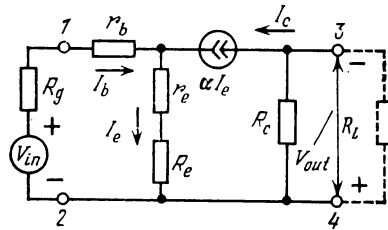


Fig. 9.5. Small-signal circuit model for a low-frequency amplifying stage

the time being to the range of rather low frequencies, we shall take α as an actual quantity and neglect the collector capacitance along with the collector junction resistance r_c , since the inclusion of the latter does not involve substantial corrections in the results of the analysis. The equivalent circuit of the amplifier stage will then be such as shown in Fig. 9.5. The circuit model also includes the internal resistance R_g of a signal generator (signal source) and shows the current and voltage increments without the sign Δ for simplicity.

From Fig. 9.5 it follows that

$$V_{in} = I_b (R_g + r_b) + I_e (R_e + r_e)$$

Substituting $I_b = (1 - \alpha) I_e$ readily gives

$$I_e = \frac{V_{in}}{R_e + r_e + (1 - \alpha) (R_g + r_b)} \quad (9.4)$$

Considering the condition (9.3), it is safe to set $I_e = V_{in}/R_e$ without introducing a large error. Knowing the current I_e , it is easy to determine all other currents and voltages in the circuit.

The coefficients, both dimensional and nondimensional, which interrelate ac components and an input signal are known as *differential parameters* of an amplifier.

Of these, the main parameter called the *amplification factor*, or *voltage gain*, is expressed through the ratio between the output and the input signal:

$$K = V_{out}/V_{in}$$

The output signal is customarily taken as an **ac component** of the collector voltage, ΔV_c , for which reason the collector potential in Fig. 9.3 is written as $V_c^0 + V_{out}$. From Fig. 9.5, it is apparent that $V_{out} = -\alpha I_e R_c$. Substituting the current I_e given by Eq. (9.4) and dividing both sides of the expression by V_{in} , we find the voltage gain in the general form

$$K = -\frac{\alpha R_c}{R_e + r_e + (1-\alpha)(R_g + r_b)} \quad (9.5a)$$

Disregarding, according to Eq. (9.3), the two last terms in the denominator, the expression reduces to a simpler form, quite suitable for all practical calculations:

$$K = -\alpha (R_c/R_e) \quad (9.5b)$$

The minus sign indicates that the polarities of the output and the input signal are different, or (in the case of a sinusoidal signal) the output is 180° out of phase with the input.

As clear from Eq. (9.5b), it is desirable that the resistance R_c can be large and R_e small. In practical circuits, however, the resistance R_c is dependent on the supply voltage and the operating point of a transistor, as obvious from (9.2b), and the resistance R_e must satisfy the stability condition (9.3). This explains why the voltage gain of the stage under study does not exceed 4 or 5.

The limitation on the voltage gain becomes expressly obvious if we transform Eq. (9.5a) by substituting R_c and R_e given in (9.2) into Eq. (9.5a). Setting $r_e \ll R_e$ then gives

$$K = -\frac{E_c - V_c^0}{E_e - V^*} \quad (9.6)$$

Let us set $E_c = 12$ V, $E_e = 3$ V, and $V_c^0 = 2$ V. At these values, $K \approx 4.5$. It is useful to remember that the voltage gain K does not depend on operating currents and approaches unity if supply voltages E_c and E_e are equal.

If an **external load** R_l shown by a dash line in Fig. 9.5 is connected to the amplifier output, then R_c in Eqs. (9.5) should be replaced by the equivalent resistance $R_c \parallel R_l$, where \parallel is a symbol identifying parallel connection.

The next important parameter of an amplifier is an *input resistance* expressed as

$$R_{in} = V_{in}/I_{in}$$

where I_{in} is the ac component of the base current. The voltage V_{in} is assumed to be applied **directly** to the base. Hence, in calculating the input resistance, one needs to set $R_g = 0$.

The input resistance plays the role of a load with respect to the signal source. So, the larger this resistance, the smaller the load on the signal source and the better the signal transfer to the stage input.

Assuming $R_g = 0$, from Fig. 9.5 we get:

$$V_{in} = I_b r_b + I_e (R_e + r_e)$$

Substituting $I_e = (\beta + 1) I_b$ and dividing both sides of the expression by $I_b = I_{in}$, we find the input resistance in the general form

$$R_{in} = r_b + (\beta + 1) (R_e + r_e) \quad (9.7a)$$

The resistances r_b and r_e may practically be neglected. So,

$$R_{in} \approx (\beta + 1) R_e \quad (9.7b)$$

If $\beta = 100$ and $R_e = 1 \text{ k}\Omega$, then $R_{in} \approx 100 \text{ k}\Omega$. Note that as R_e grows, the input resistance cannot rise infinitely as obvious from Eqs. (9.7). The causes of the limit are discussed in Sec. 9.4.

The third important parameter of an amplifier is an *output resistance* given by

$$R_{out} = \frac{(V_{out})_{oc}}{I_{out\ sc}}$$

where $(V_{out})_{oc}$ is the output voltage in the open-circuit (no-load) condition of the stage, that is, in the absence of the external resistor R_l ; and $I_{out\ sc}$ is the output current in the short-circuit condition (what is meant here is short-circuiting for ac components).

The output resistance characterizes the **load capacity** of the stage: the lower this resistance, the larger the current that can be delivered to the external load and so the smaller the external resistance that can be inserted into the circuit.

From the physical viewpoint, the output resistance of a circuit is an incremental resistance that can be measured at the output terminals in the absence of an input signal ($V_{in} = 0$) with the external load disconnected ($R_l = \infty$). The theoretical calculations of R_{out} are performed under the same conditions too.

For the circuit of Fig. 9.5, there is no need to carry out special calculations. Since the input of the circuit is separated from the output by a current generator, which is idle at $V_{in} = 0$, we may just write:

$$R_{out} = R_c \quad (9.8)$$

The inclusion of the collector junction resistance (that is, the current generator internal resistance) does not practically affect the above equality.

9.3.3. Drift of dc components. Under the quiescent conditions (at $V_{in} = 0$), the currents and voltages of an amplifier can vary with temperature, supply voltage, and under the action of other factors. The slowly varying **uncontrollable** increments in currents and voltages that result from the above factors are known as the *drift of dc components*.

If an input signal changes rather rapidly (that is, at a sufficiently high frequency), then it is an easy problem to distinguish the amplified signal from the drift. But in dc amplifiers intended to amplify slowly varying signals, the drift (**parasitic increments**) is generally indistinguishable from the increments induced by the useful signal. So, the value of drift places a limit on the level of sensitivity of a dc amplifier, that is, on its capability of amplifying small signals.

Because dc amplifiers used in microelectronics occupy a leading place, the drift problem proves particularly acute.

Let uncontrollable processes have brought about the increments ΔE_c and ΔE_e and also the increments ΔV^* and $\Delta \alpha$ of the transistor parameters. From Eq. (9.2b) it then follows that the collector potential V_c^0 changes by

$$\Delta V_c^0 = \Delta E_c + \Delta I_c^0 R_c$$

Here, the plus sign is due to the fact that the increments ΔE_c and $\Delta I_c^0 R_c$ are independent, or *uncorrelated*, and hence can add together in the worst case. Allowing for the relation $I_c^0 = \alpha I_e^0$, the collector current change may be written in the form

$$\Delta I_c^0 = \Delta \alpha I_e^0 + \alpha \Delta I_e^0$$

Last, using Eq. (9.2a), the emitter current change may be given by

$$\Delta I_e^0 = \frac{\Delta E_e + \Delta V^* + \Delta \alpha I_e^0 R_b}{R_e + (1 - \alpha) R_b}$$

Making appropriate substitutions and dividing ΔV_c^0 by the voltage gain of Eq. (9.5a), we obtain the voltage called the *referred drift*:

$$V_{dr} = \frac{\Delta V_c^0}{|K|}$$

This quantity is convenient for use in estimating the smallest input signal yet distinguishable against the background of drift.

If we disregard a small difference between the denominators of Eqs. (9.2a) and (9.5a) and replace α by a more convenient factor B, the referred drift will take the form

$$V_{dr} = \Delta V^* + \left(\Delta E_e + \frac{\Delta E_c}{|K|} \right) + \frac{\Delta B}{B} \frac{I_e^0 (R_e + R_b)}{B + 1} \quad (9.9)$$

This expression says that among the four causes of drift (ΔB , ΔE_e , ΔE_c , and ΔV^*), the component ΔV^* is **principally unavoidable**. The remaining components can generally be reduced by a circuit design including the stabilization of supply voltage and a decrease in operating current.

To estimate the drift, let us set $B = 100$, $\Delta B/B = 0.5$, $I_e^0 = 1$ mA, $R_e + R_b = 3$ k Ω , $E_e = 3$ V, $E_c = 12$ V, and $|K| = 5$.

The third component of the drift will then be equal to 15 mV. To decrease ΔE_e and ΔE_c , we should stabilize the supply voltage to 0.1%. As for the unavoidable component ΔV^* , its value primarily depends on the range of temperature variations. Assuming the temperature sensitivity ε to be equal to $1.5 \text{ mV}^\circ \text{C}^{-1}$ (see p. 90) and setting $\Delta T = 100^\circ \text{C}$, the component ΔV^* reaches 150 mV. This component is obviously the largest. As clear from the above example, the referred drift sets a lower limit on input signals of the order of 100 mV.

9.3.4. Cascading. The voltage gain of a simple single-stage amplifier has a limit and does not usually exceed 4 or 5. Therefore, to obtain the desired gain that often has to range into a few thousands and even into the tens of thousands, it is necessary to connect several similar amplifying stages in series to form a multistage (cascade) amplifier. In a cascaded configuration, the output of each stage is connected to the input of the next stage. The overall gain will be the product of the gains of all the stages:

$$K = K_1 K_2 \dots K_n \quad (9.10a)$$

or, if the stages are identical,

$$K = K_1^n \quad (9.10b)$$

Here, n is the number of stages, and K_1 is the gain of an individual stage.

The arrangement of individual stages in a series circuit is known as *cascading* (*cascade connection*).

From Eq. (9.10b) it is easy to find the required number of stages for the specified values of K and K_1 :

$$n = \log K / \log K_1$$

If $K = 10^4$ and $K_1 = 4$, then $n \approx 7$. As seen, this type of amplifier can consist of a rather large number of stages.

In cascading a simple amplifier, one has to bring under control specific problems which add to the difficulty of obtaining high gains. We shall not dwell on these problems since the simplest amplifiers discussed above are not popular in microelectronics.

9.4. Transients in a Simple Amplifier

Externally the transistor in a simple amplifier is connected in a common-emitter (CE) configuration: an input signal is applied to the base and the output signal is taken off the collector. But *in essence the transistor operates in a common-base (CB) connection*, since both in the quiescent state and with an input signal applied, the specified

value is found to be the emitter current rather than the base current. For an ac component, this conclusion is evident from Eq. (9.4) considering the condition (9.3).

Hence, applying a signal V_{in} to the input, we thus set a step of emitter current, ΔI_e . Correspondingly, the transient characteristics of the collector current and collector voltage will be determined by the transient response of gain α . Considering the effect of collector capacitance, it is necessary to use $\tau_{\alpha oe}$ of Eq. (4.66) rather than τ_α . Substituting the transform α (s) into Eq. (9.5b), we find the Laplace transform K (s). Proceeding from the latter quantity, it is then easy to find the transient and frequency characteristics. In a relative scale, all these characteristics will coincide with the appropriate characteristics for gain α (see Sec. 4.7). So there is no need to give again the corresponding formulas and graphs.

Instead, we shall consider the case where the emitter current cannot be regarded as a specified value, that is, when the summand $(1-\alpha)(R_g + r_b)$ in the denominator of Eq. (9.5a) is comparable with the first two summands. This case is not typical for a simple amplifier, but is of importance from the methodical standpoint since it serves as the basis for the analysis of other types of amplifier.

9.4.1. Time constant at high frequencies. The equivalent circuit of an amplifier, intended for work in the range of high frequencies

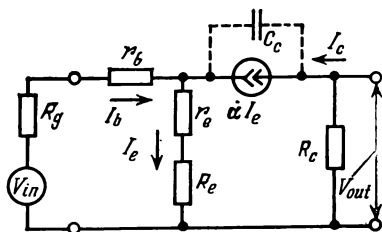


Fig. 9.6. Small-signal circuit model for a high-frequency amplifying stage

(pulses of a short length) appears in Fig. 9.6. We shall disregard the collector capacitance C_c in the circuit but take into account its effect (as we did in Sec. 4.7) through the equivalent time constant

$$\tau_{\alpha oe} = \tau_\alpha + C_c R_c \quad (9.11)$$

Let us use the simplest transform (4.48) for α and substitute it into Eq. (9.5a). Multiplying the numerator and denominator by the binomial $1 + s\tau_{\alpha oe}$ and reducing the denominator to the form $A(1 + sa)$, we can write the Laplace transform of the stage gain:

$$K(s) = K/(1 + s\tau_{hf}) \quad (9.12)$$

Here τ_{hf} is the time constant of the stage at high frequencies (short pulse lengths). This quantity may be easily reduced to the form

$$\tau_{hf} = \tau_{\alpha oe} / (1 - \alpha \gamma_e) \quad (9.13)$$

where γ_e is current dividing coefficient for the input section:

$$\gamma_e = (R_g + r_b) / (R_e + r_e + R_g + r_b) \quad (9.14)$$

The coefficient γ_e defines the fraction of current αI_e that branches off from the current source into the emitter circuit during the growth of α (at the start, $\alpha = 0$).

If $R_g + r_b \ll R_e + r_e$, as was the case so far, then $\gamma_e \approx 0$. This means that the current αI_e does not practically branch out into the emitter circuit: almost all the current αI_e goes into the low-resistance base circuit. The original step of emitter current, $I_e(0)$, thus remains invariable, and the transient proceeds with a time constant $\tau_{hf} = \tau_{\alpha oe}$, as we have mentioned earlier in the beginning of this section.

A noticeable rise in τ_{hf} as against $\tau_{\alpha oe}$ begins only when the resistances in the emitter and base circuits become comparable in value. In the limit, if $R_g + r_b \gg R_e + r_e$, the coefficient γ_e approaches unity. So the time constant τ_{hf} becomes equal to $\tau_{\alpha oe} / (1 - \alpha)$ which is the equivalent time constant of the CE configuration, as follows from Eq. (4.67):

$$\tau_{oe} = \tau + (\beta + 1) C_c R_c \quad (9.15)$$

Such a transformation of τ_{hf} is quite natural since the condition $\gamma_e = 1$ means that all the current αI_e flows to the emitter circuit, and hence *the base current remains invariable*, specified. The transistor should thus be considered connected in a CE configuration which, as known, is characterized by a time constant τ_{oe} .

Note that we could express τ_{hf} through τ_{oe} from the very beginning. For this, it is enough to replace α in Eq. (9.13) by $\beta / (\beta + 1)$ and multiply the numerator and denominator by $\beta + 1$. After making elementary transformations and taking into account Eq. (4.55), we get

$$\tau_{hf} = \tau_{oe} / (1 + \beta \gamma_b) \quad (9.16)$$

Here γ_b is the current dividing coefficient defining the amount of current αI_e that flows to the base circuit:

$$\gamma_b = 1 - \gamma_e = \frac{R_e + r_e}{R_e + r_e + R_g + r_b} \quad (9.17)$$

As γ_b rises, τ_{hf} decreases and, in the limit, when $\gamma_b = 1$, reaches a minimum value of $\tau_{\alpha oe}$.

From the above it follows that *the speed of response of a simple amplifying stage rises* (that is, τ_{hf} becomes smaller) *if the resistance*

in the base circuit is at minimum and that in the emitter circuit at maximum. Hence, the lower the signal source resistance, the better the transient response of the amplifying stage.

9.4.2. Transient and frequency characteristics. Multiplying the current gain transform of Eq. (9.12) by V_{in} gives the transform of an output voltage. The original of this transform will be a transient response

$$V_{out}(t) = KV_{in}(1 - e^{-t/\tau_{hf}}) \quad (9.18)$$

shown in Fig. 9.7 in two variants. The first is typical for a low-resistance signal source, where the specified value is in fact the emitter current, that is, $\gamma_e \approx 0$ and $\tau_{hf} \approx \tau_{\alpha oe}$ (see Fig. 9.7a);

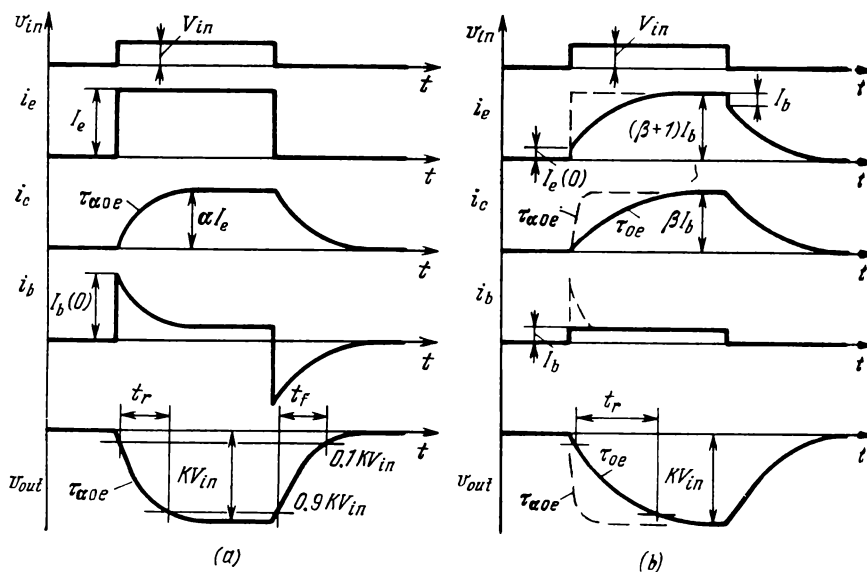


Fig. 9.7. Transients in a stage

(a) at specified emitter current; (b) at specified base current

the second is typical for a high-resistance signal source, where the specified value is the base current, that is, $\gamma_e \approx 1$ and $\tau_{hf} \approx \tau_{oe}$ (see Fig. 9.7b). For the illustrative purpose, the steady-state values of currents and output voltages are taken equal. In Fig. 9.7b are also shown dash curves corresponding to the specified emitter current (see Fig. 9.7a).

After the input signal ceases, the voltage V_{out} drops to zero with the same time constant τ_{hf} , therefore both slopes of the output signal

are equal. Counting off the rise time and the fall time between the 10% and 90% levels of the steady-state values of KV_{tn} gives

$$t_r = t_f = 2.2\tau_{hf} \quad (9.19)$$

Setting $\tau_{\alpha oe} = 5$ ns and $\gamma_e = 0.4$, we have $\tau_{hf} \approx 8$ ns and $t_r = t_f \approx 17.5$ ns.

Frequency response for the gain results after replacing the Laplace variable s in the transform (9.12) by $j\omega$:

$$\dot{K} = \frac{K}{1 + j\omega/\omega_{hf}} \quad (9.20)$$

where $\omega_{hf} = 1/\tau_{hf}$ is the cutoff angular frequency for the gain at a level of 0.7, or 3 dB.

The modulus and phase of the complex gain \dot{K} represent respectively the amplitude-frequency and phase-frequency response:

$$K(\omega) = \frac{K}{\sqrt{1 + (\omega/\omega_{hf})^2}} \quad (9.21a)$$

$$\varphi(\omega) = -\arctan(\omega/\omega_{hf}) \quad (9.21b)$$

The shape of these characteristics is the same as for α (see Fig. 4.21).

Assume that $\tau_{\alpha oe} = 5$ ns and $\gamma_e = 0.4$, as we did in the preceding example. Then, $f_{hf} = (1/2\pi)\omega_{hf} \approx 20$ MHz.

The input base current that determines the input resistance of an amplifier deserves particular consideration. Immediately after applying an input signal, when $i_c = 0$, the base current is equal to the emitter current (see Fig. 9.7a):

$$I_b(0) = I_e$$

The current $I_b(0)$ is $\beta + 1$ times as high as the steady-state value. Correspondingly, the initial input resistance is by a factor of $\beta + 1$ below the steady-state value given by (9.7). The base current then drops with time, while the input resistance rises until it reaches the steady-state value. But since the steady-state base current is only $1/\beta$ as large as the collector current, the transient of base current proves β times longer, approximately βt_r .

In conclusion, let us point out that in the presence of external load resistance R_l the expressions for $\tau_{\alpha oe}$ and τ_{oe} should include a smaller quantity $R_c \parallel R_l$ rather than R_c .

9.5. Simple MOSFET Amplifiers

In use are two circuit versions of these amplifiers, with a resistive load and with a dynamic load (Fig. 9.8). In amplifying stages, MOSTs always operate in the flat regions of the characteristics, where the transconductance and gain of transistors are the highest.

9.5.1. Resistive-load amplifier. In this amplifier illustrated in Fig. 9.8a, the quiescent state is characterized by the following potentials:

$$V_s^0 = -E_s \quad (9.22a)$$

$$V_d^0 = E_d - I_d^0 R_d \quad (9.22b)$$

To bias a transistor into conduction, the voltage V_{gs}^0 must be in excess of the threshold voltage; hence, in this circuit the condition $E_s > V_0$ must be met. It is advisable to get the potential V_d^0 to be

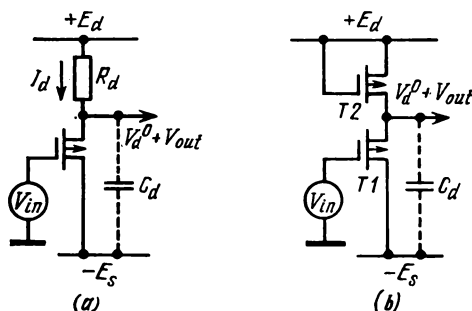


Fig. 9.8. MOS transistor amplifiers
(a) resistive-load; (b) dynamic load

equal to zero. This facilitates cascading of amplifiers, namely, permits connecting drain of the preceding stage directly to the gate of the next stage.

The quiescent current I_d^0 can be easily found by substituting $V_{gs}^0 = V_g^0 - V_s^0$ in Eq. (5.8):

$$I_d^0 = 1/2 b (E_s - V_0)^2 \quad (9.23)$$

whence, setting the current I_d^0 , we readily determine the required value of E_s .

If supply voltages E_d and E_s are constant, the drift of dc components I_d^0 and V_d^0 is due primarily to the drift of parameters V_0 and b . As known (see p. 163), there is a critical value of current I_d at which its temperature drift is minimum (over the narrow range of temperatures this drift is close to zero). At currents above the critical, the TC of current is positive, and at currents below the critical, the TC is negative.

We now turn to the estimation of voltage gain. Taking the incremental resistance of a drain in the flat region to be at infinity ($r_d = \infty$), from the small-signal circuit model of Fig. 9.9 it follows that $I_d = S V_{in}$, and hence

$$V_{out} = -I_d R_d = -S R_d V_{in}$$

The voltage gain then assumes the form

$$K = V_{out}/V_{in} = -SR_d \quad (9.24)$$

If the resistance r_d has a finite value comparable to R_d , the total current SV_{in} will be distributed between the branches R_d and r_d . The drain current becomes equal to

$$I_d = SV_{in}r_d/(r_d + R_d)$$

Correspondingly,

$$V_{out} = -I_d R_d = -S (r_d \parallel R_d) V_{in}$$

The voltage gain may then be written in the form

$$K = -\frac{\mu}{1 + r_d/R_d} \quad (9.25)$$

where $\mu = Sr_d$ is the amplification factor of a transistor [see Eq. (5.18) and footnote on p. 160].

From Eq. (9.25) it is obvious that the voltage gain can reach its maximum ($|K| = \mu$) if $R_d \gg r_d$. This condition is practically

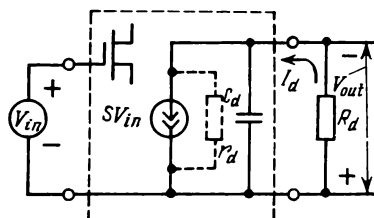


Fig. 9.9. Amplifying stage drain-circuit model

impossible to meet because a voltage drop $I_d^0 R_d$ turns out to be large, and so there is a need for a high supply voltage E_d [see Eq. (9.22b)]. The resistance R_d , therefore, is equal to or smaller than 0.2 or $0.3r_d$, and hence $|K| \leq 0.2\mu$. It can easily be shown that Eq. (9.24) applies in this case.

9.5.2. Dynamic-load amplifier. In this amplifier shown in Fig. 9.8b, the load transistor $T2$ operates in the flat region of the characteristic. Therefore, the resistance the transistor offers to small signals can be found by differentiating the current I_{d2} with respect to the voltage V_{ds2} entering into Eq. (8.47). Considering Eq. (5.19a), we obtain

$$R_d = dV_{ds2}/dI_{d2} = 1/S_2 \quad (9.26)$$

where S_2 is the transconductance of $T2$. The internal resistance r_{d2} here is taken infinite; its inclusion usually is of no consequence.

Replacing R_d by $1/S_2$ in Eq. (9.25), μ by μ_1 , and substituting $r_d = \mu_1/S_1$, we find

$$K = -\frac{\mu_1}{\mu_1(S_2/S_1) + 1} \approx -\frac{S_1}{S_2}$$

The inequality $\mu(S_2/S_1) \gg 1$ used to simplify the expression is substantiated below.

Since the currents in both transistors are equal, the ratio S_1/S_2 , according to Eq. (5.19b) may be written in the form

$$\frac{S_1}{S_2} = \sqrt{\frac{b_1}{b_2}}$$

Let us introduce the factor B which, with regard to Eq. (5.7), characterizes the geometry of transistors:

$$B = \frac{b_1}{b_2} = \frac{Z_1/L_1}{Z_2/L_2} \quad (9.27a)$$

With the lengths of channels being equal,

$$B = Z_1/Z_2 \quad (9.27b)$$

The voltage gain may thus be written in the form

$$K = -\sqrt{B} \quad (9.28)$$

So, the voltage gain is determined by the dimensions of channels of the active and load transistors, first of all, by the ratio of the widths of channels. The ratio Z_1/Z_2 higher than 50 to 100 is difficult to achieve, therefore the voltage gain as a rule reaches merely a few units.

Note that the voltage gain is associated with the quiescent condition of an amplifier. Indeed, equating the currents of both transistors and using the factor B , it is easy to obtain the relation

$$\frac{V_{gs2}^0 - V_0}{V_{gs1}^0 - V_0} = \sqrt{B} = |K|$$

Substitute here the values of $V_{gs1}^0 = E_s$ and $V_{gs2}^0 = E_d - V_d$, which are apparent from Fig. 9.8b. The relation between the voltage gain and the quiescent condition then assumes the form

$$\frac{E_d - (V_d^0 + V_0)}{E_s - V_0} = |K| \quad (9.29)$$

This expression, both in structure and in essence, is analogous to expression (9.6) for a bipolar amplifier. The voltage E_s must obviously be much smaller than E_d but noticeably higher than V_0 to exclude instability.

9.5.3. Transients. In MOS transistor amplifiers, transients are associated with the recharge of parasitic capacitance C_d connected to the drain of the active transistor (as shown by a dash line in Fig. 9.8). This capacitance is the same in structure as that given by Eq. (8.52) for MOS transistor switches.

Assume a step of current arrives at the input of the amplifier. The drain current will then change practically at an instant (with a very small time constant τ_s), but the drain voltage will change exponentially with a time constant τ_c .

The Laplace transform of the voltage gain is easy to obtain from Eq. (9.24) replacing R_d by the impedance

$$Z_d = R_d \parallel \frac{1}{sC_d}$$

After elementary transformations, the voltage gain reduces to the form

$$K(s) = \frac{K}{1 + s\tau_c} \quad (9.30)$$

where $\tau_c = C_d R_d$. For a dynamic-load amplifier, R_d is understood to be the incremental resistance $1/S_2$ of Eq. (9.26).

The original of the transform given by (Eq. 9.30) is the simplest exponential function we have dealt with more than once [see, for example, Eq. (9.18)]. The expressions such as (9.19) can apply to determine the rise time and fall time.

The expression for the complex voltage gain also has a traditional structure resulting from Eq. (9.30)

$$\dot{K} = \frac{K}{1 + j\omega/\omega_c} \quad (9.31)$$

where $\omega_c = 1/\tau_c$.

Setting $R_d = 20$ k Ω , $C_d = 3$ pF, we have $\tau_c = 60$ ns, $t_r = 130$ ns, and $f_c \approx 2.7$ MHz.

9.5.4. Miller effect. In Subsec. 8.7.4 we mentioned this effect in describing Eq. (8.52) for the total capacitance of a MOSFET.

The Miller effect shows up in that the equivalent input admittance of an active two-port (fourpole), being attributed to feedback, differs from the admittance inserted in the feedback circuit.

Consider a concrete circuit of Fig. 9.10a having a complex admittance Y connected between the output (drain) and input (gate). Apply an ac voltage component V to the input. The voltage derived from the output then becomes KV , where K is the voltage gain. The potential difference obtained across Y thus has the form

$$V - KV = V(1 - K)$$

This potential difference causes a current

$$I = V(1 - K)Y$$

Since the current is taken off the input signal source, the ratio I/V is the *equivalent input admittance* Y_{eq} shown in Fig. 9.10b. Its expression is $Y_{eq} = Y(1 - K)$.

In the circuit under discussion and in most of the other circuits, the voltage gain is negative. We then can use a more convenient form

$$Y_{eq} = (|K| + 1)Y$$

As obvious, when $|K| \gg 1$ the equivalent admittance can be much higher than the admittance really connected in the circuit.

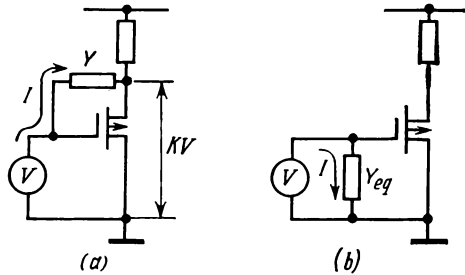


Fig. 9.10. Miller effect

(a) real circuit with admittance; (b) circuit with equivalent admittance

In practice, the admittance Y most often represents a capacitance (C_{gd} in the given case). The Miller effect, therefore, commonly involves an apparent increase in the input capacitance:

$$C_{eq} = (|K| + 1)C \quad (9.32)$$

[in Eq. (8.32) the quantity K stands for $|K| + 1$]. However, if the circuit gain is positive but lower than unity (as in the follower, see Sec. 9.7), the equivalent input capacitance has the form $(1 - K)C$ and proves much smaller than the real capacitance between input and output, C_{gd} .

Let us point out that the Miller effect has a general meaning: it applies to the analysis of other devices apart from single-stage amplifier circuits and the type of active devices.

9.6. Differential Amplifiers

A differential amplifier (DA) circuit shown in Fig. 9.11 consists of two identical (symmetric, or balanced) branches, each containing a transistor and resistor. The current source I_0 is connected to the

common emitter circuit. The output voltage is the difference between collector potentials, and the input voltage is the difference between base potentials.

The structure of a DA is the same as for the current switch of Fig. 8.13, but the mode of operation is different: neither of the transistors is in the off condition and both work in the active mode. The use of current source I_0 ensures the stability of a steady-state point (that is, of currents I_c and voltages V_s).

9.6.1. Principle of action. What underlies the function of a differential amplifier is the ideal symmetry of both of its branches, that is, identity of the parameters of transistors $T1$, $T2$ and equality of resistances R_{c1} , R_{c2} . In the absence of a signal, the currents and collector potentials will be equal, and the output voltage will be zero. Because of symmetry, V_{out} remains at zero with the simultaneous and equal changes of currents in both branches, whatever the causes of these changes. Hence, *in an ideal DA the drift of output voltage does not exist*, though in each of the branches the drift can be comparatively large.

Let us apply equal base voltages ($\Delta V_{b1} = \Delta V_{b2}$) known as common-mode (in-phase) signals. These signals cause the emitter potentials to change by the same amount as that for base potentials: $\Delta V_e = \Delta V_b$ (since the emitter junction voltages V^* may be considered invariable). In the case of an ideal current source I_0 (where $R_i = \infty$), the increment ΔV_e does not tend to alter currents in the DA branches. The collector potentials do not change and the output voltage remains equal to zero. If $R_i \neq \infty$, an increment ΔI_0 appears: but it will be equally distributed between the two branches, so the changes in collector potentials will also be equal. Thus, in this case too, $V_{out} = 0$. This means that *in an ideal DA, common-mode signals have no effect on the output voltage*.

Apply now base voltages equal in magnitude but opposite in sign ($\Delta V_{b1} = -\Delta V_{b2}$), known as *differential* signals. By definition, the difference between these signals is an input signal for the amplifier:

$$V_{in} = \Delta V_{b1} - \Delta V_{b2}$$

On the strength of symmetry, the input signal V_{in} will be equally divided between both emitter junctions: at one of the junctions the

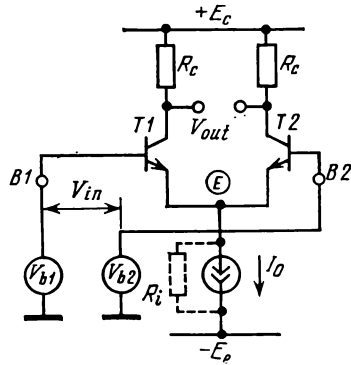


Fig. 9.11. Differential amplifier

voltage V^* will rise by $1/2 V_{in}$ and at the other the voltage will drop by the same amount. The increments in currents and collector potentials in the DA branches will thus be equal but opposite in sign. The difference-mode operation of the amplifier provides an output voltage

$$V_{out} = \Delta V_{c1} - \Delta V_{c2}$$

As clear, an ideal DA responds only to a differential signal, hence the name of this type of amplifier.

Since a differential signal is equally divided between the emitter junctions, the midpoint potential (emitter potential) remains in-

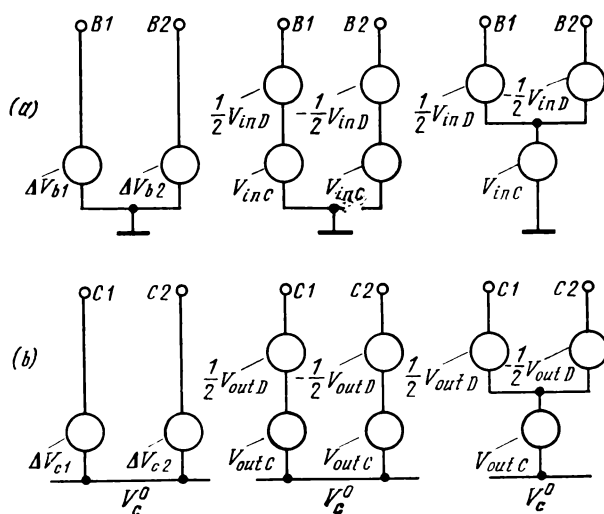


Fig. 9.12. Common-mode and difference-mode components of an input (a) and an output (b) signal

variable. So, in the analysis of differential signals the potential V_c can be regarded to be specified and the point E grounded for ac components.

Any combination of voltages ΔV_{b1} and ΔV_{b2} may be represented as a sum of the common-mode and the difference-mode component (Fig. 9.12a):

$$\Delta V_{b1} = V_{in\ c} + 1/2 V_{in\ D} \quad (9.33a)$$

$$\Delta V_{b2} = V_{in\ c} - 1/2 V_{in\ D} \quad (9.33b)$$

This representation is convenient because it permits analyzing each of the components separately.

The common-mode and the difference-mode component of an input signal are expressed in the following manner:

$$V_{in\ C} = 1/2 (\Delta V_{b1} + \Delta V_{b2}) \quad (9.34a)$$

$$V_{in\ D} = \Delta V_{b1} - \Delta V_{b2} \quad (9.34b)$$

The output voltage can also be represented by the sum of common-mode and difference-mode components:

$$V_{out\ C} = 1/2 (\Delta V_{c1} + \Delta V_{c2}) \quad (9.35a)$$

$$V_{out\ D} = \Delta V_{c1} - \Delta V_{c2} \quad (9.35b)$$

where ΔV_c is the increment in collector potentials *with respect to the steady-state potential* V_c^0 . In Fig. 9.11, the difference components $V_{in\ D}$ and $V_{out\ D}$ do not have the subscript "D".

It is highly important for the operation of a DA that its current I_0 be invariable. If the current source is ideal ($R_i = \infty$, see Fig. 9.11), the common-mode component of a signal produces only an increment in the emitter potential: $\Delta V_e = \Delta V_b = V_{in\ C}$. The currents in the branches and collector potentials remain invariable.

If the current source is not ideal and thus has a finite resistance R_i , the increment ΔV_e causes an increment $\Delta I_0 = \Delta V_e / R_i$. This increment, being distributed between both branches of the amplifier, is responsible for collector potential increments ΔV_{c1} and ΔV_{c2} . If the circuit branches are identical, these increments are equal, $\Delta V_{c1} = \Delta V_{c2}$. So, only the common component will appear at the output as clear from Eqs. (9.35). Where the circuit branches are nonidentical, the increments in collector potentials are different: $\Delta V_{c1} \neq \Delta V_{c2}$. The output voltage will then contain a **parasitic** difference component along with the common-mode component.

Since in an ideal DA a common-mode input signal should not cause a common-mode output signal, let alone a differential signal, the current-source internal resistance should meet most stringent requirements.

That the operation of a DA relies on the identity of its branches explains why these amplifiers (and DA-based circuits) are so popular in microelectronics. Only in ICs, where the elements are tens of micrometers distant from each other, is it possible to ensure the identity of parameters, TCs, and other quantities. Besides, the number of elements that aid in improving the quality of a micro-circuit is not critical.

9.6.2. Gains. In a practical differential amplifier, where the branches are nonidentical and a current source has a finite resistance, the common-mode component of an input signal affects the difference component of an output signal, while the input-signal difference component affects the output-signal common-mode component.

In the general case, the relations between the common-mode and difference-mode components can be expressed as

$$V_{out\ C} = K_{CC}V_{in\ C} + K_{CD}V_{in\ D} \quad (9.36a)$$

$$V_{out\ D} = K_{DC}V_{in\ C} + K_{DD}V_{in\ D} \quad (9.36b)$$

Here the factors K are voltage gains of respective voltage components from input to output. In an ideal DA, mutual gains K_{CD} and K_{DC} are equal to zero.

Consider the main parameter of a DA—the *difference-component gain* K_{DD} , which is often called just the *voltage gain* and designated as K .

As noted earlier in the preceding section, the emitter potential remains unchanged on applying a differential signal and, hence, should be taken equal to zero for ac components. Therefore, the gain for each circuit branch can be obtained from Eq. (9.5a), setting $R_e = 0$. Since each of the branches amplifies the signal $1/2V_{in}$ and the amplified signals add up at the output, the gain of a DA is equal to the gain of the individual branch.

Assuming $R_e = 0$, from Eq. (9.5a) we obtain

$$K = \frac{\alpha R_c}{r_e + (1 - \alpha)(R_g + r_b)} \quad (9.37a)$$

Obviously, the gain of a differential amplifier is much higher than for a simple amplifier and can be tens of times as high. So a differential amplifier, apart from being free of drift (or having a rather small value of drift), *produces a fairly high gain*, which is its second important merit.

In the case of low-resistance signal sources (R_g being smaller than 1 k Ω) and low working currents (less than 1 mA), the second term in the denominator of Eq. (9.37a) can be neglected. Hence,

$$K = -\alpha (R_c/r_e) \quad (9.37b)$$

Substituting here R_c from Eq. (9.2b) and $r_e = \varphi_T/I_e^0$ from (4.41) gives the gain of the form

$$K = -(E_c - V_c^0)/\varphi_T \quad (9.38)$$

Assuming $E_c = 12$ V and $V_c^0 = 2$ V, we have $K = -400$. As with a simple amplifier, the gain of a DA is seen to be related to the supply voltage and the quiescent collector voltage [cf. Eq. (9.6) and (9.29)]. But the denominator in Eq. (9.38) has a much smaller value, unattainable in simple amplifiers because of stability requirements. As with other circuits discussed above, the gain (at a given V_c^0) is independent of working current. It depends on temperature through the quantity φ_T .

The common-mode component gain, according to Eqs. (9.36), is defined as

$$K_{CC} = \frac{V_{out\ C}}{V_{in\ C}} \Big|_{v_{in\ D}=0}$$

In Fig. 9.11, therefore, we need to connect the two bases together, with the signal $V_{in\ C}$ applied to both. Setting $V^* = \text{constant}$ gives $\Delta V_e = V_{in\ C}$. If the current source resistance here is equal to R_t , the current I_0 changes by $\Delta I_0 = V_{in\ C}/R_t$, and collector potentials by $-\alpha (1/2\Delta I_0) R_c$. Then

$$K_{CC} = - (\alpha R_c)/2R_t \quad (9.39)$$

Commonly, $R_c/R_t < 1$, and hence $K_{CC} < 1$.

The gain K_{CD} , according to Eqs. (9.36), characterizes the effect of the difference input component on the common-mode component of output voltage:

$$K_{CD} = \frac{V_{out\ C}}{V_{in\ D}} \Big|_{v_{in\ C}=0}$$

Since the differential signal is equally distributed between the two emitter junctions, the main cause of changes in the mean collector potential is unbalance of the gains of circuit branches. We thus can assume

$$K_{CD} = \Delta K$$

where $\Delta K = K_1 - K_2$. Multiplying and dividing the right side by the mean gain

$$K = 1/2 (K_1 + K_2)$$

and considering that the main cause of unbalance of the gains is the difference between the resistances R_c , we set $\Delta K/K = \Delta R_c/R_c$. So,

$$K_{CD} = K (\Delta R_c/R_c) \quad (9.40)$$

For example, if $(\Delta R_c/R_c) = 0.02$, then $K_{CD} = 0.02 K$. Hence, a change in the dc component of collector potentials due to a differential signal is a few orders of magnitude smaller than the output voltage.

9.6.3. Common-mode rejection ratio. According to Eq. (9.36), the gain K_{DC} defines the effect of the common component of an input signal on the difference component of an output signal. This effect is rather substantial because the summand $K_{DC}V_{in\ C}$, being indistinguishable from the summand $K_{DD}V_{in\ D}$, is equivalent to a false signal. Since in practice the component $V_{in\ C}$ can be thousands of times as large as $V_{in\ D}$, the value of K_{DC} must be smaller than K_{DD} by a few orders of magnitude. The ratio between the absolute values of these two quantities, is defined as the *common-mode rejection*

ratio (CMRR) expressed in decibels and denoted here as K_R :

$$K_R = 20 \log \left| \frac{K_{DD}}{K_{DC}} \right| \quad (9.41)$$

Thus, if $|K_{DD}/K_{DC}| = 10^4$, then $K_R = 80$ dB.

A widespread type of common-mode signal is various noise (both internal and external) and also stray pickup, which act simultaneously on both inputs. That is why *an increase in K_R is one of the basic ways in raising the noise stability of a differential amplifier.*

To evaluate common-mode rejection, let us suppose that the current source resistance is equal to R_i (the same as in the above examples) and the asymmetry of DA branches is as follows:

$$\begin{aligned} \alpha_1 &= \alpha + \Delta\alpha, & \alpha_2 &= \alpha - \Delta\alpha \\ R_{c1} &= R_c + \Delta R_c, & R_{c2} &= R_c - \Delta R_c \end{aligned}$$

where α and R_c are mean values.

As mentioned earlier, an increment in current I_0 due to a common-mode input signal is $V_{in\ c}/R_i$. Let this increment be equally divided between the emitter junctions. The changes in collector potentials can then be written as

$$\begin{aligned} \Delta V_{c1} &= -\alpha_1 (V_{in\ c}/2R_i) R_{c1} \\ \Delta V_{c2} &= -\alpha_2 (V_{in\ c}/2R_i) R_{c2} \end{aligned}$$

Equating the difference between these increments to the augend on the right of Eq. (9.36b) and substituting the above expressions for α_1 , α_2 , R_{c1} , and R_{c2} , it is easy to define

$$K_{DC} = -\alpha \frac{R_c}{R_i} \left(\frac{\Delta\alpha}{\alpha} + \frac{\Delta R_c}{R_c} \right)$$

The common-mode rejection ratio in the linear (not logarithmic) form can be found dividing the gain K_{DD} by K_{DC} . Using Eq. (9.37b), we shall represent K_R in the following general form

$$K_R = \frac{1}{\delta} \frac{R_i}{r_e} \quad (9.42)$$

Here δ is the *asymmetry (unbalance) coefficient* for a DA, that is, the sum of relative spreads of the parameters in its branches:

$$\delta = \frac{\Delta\alpha}{\alpha} + \frac{\Delta R_c}{R_c}$$

This sum can be supplemented, of necessity, with the spread in other parameters of transistors.

In deriving Eq. (9.42) the signs of increments $\Delta\alpha$ and ΔR_c were taken positive since it was assumed that $R_{c1} > R_{c2}$ and $\alpha_1 > \alpha_2$. The asymmetry coefficient was thus found to be the arithmetic sum of relative increments. In practice, both increments are **independent**

and **uncontrollable**, that is, they can differ in value and sign. The worst-case approach presupposes the calculation of the asymmetry coefficient as the **sum of moduli** of the maximum possible or most probable relative increments. In practical DAs, the coefficient δ can be smaller and K_R larger than the calculated values.

Since they have a smaller asymmetry coefficient, IC differential amplifiers ensure larger values of K_R than DAs using discrete transistors.

From expression (9.42) an important conclusion follows: *the common-mode rejection ratio is directly dependent on the current source resistance R_i* . Consequently, this resistance must be as high as possible. The simplest current sources of the resistive type (see Fig. 8.15) are unsuitable for use in DAs.

9.6.4. Input resistance. Consider the input resistance of a DA for the difference component and that for the common component of a signal. These resistances differ substantially in value.

For a difference component, the input resistance is twice the input resistance of each half of the DA. Using Eq. (9.7a) and setting $R_e = 0$, we get

$$R_{in D} = 2 [(\beta + 1) r_e + r_b] \quad (9.43)$$

If $\beta = 100$, $r_e = 25 \Omega$, and $r_b = 150 \Omega$, then $R_{in D} = 5.35 \text{ k}\Omega$. The resistance r_e is inversely proportional to the quiescent current I_e^0 . To increase the input resistance, therefore, it is advantageous to use a DA in the region of small currents (in the microampere region). Besides, it is practicable to employ transistors with a high gain β , for example, Darlington pairs (see Sec. 9.2). Thus if $I_e^0 = 50 \mu\text{A}$ and $\beta = 2000$, then $r_e = 0.5 \text{ k}\Omega$ and $R_{in D} \approx 2 \text{ M}\Omega$.

For a common component, the input resistance is a function of the current source resistance R_i . Setting $\Delta V_e = V_{in C}$ gives the increment $\Delta I_0 = V_{in C}/R_i$. Correspondingly, $\Delta I_e = 1/2 \Delta I_0$ and $\Delta I_b = 1/2 (1 - \alpha) \Delta I_0$. Dividing $V_{in C}$ by $2\Delta I_b$ and passing from α to β , we find the common-mode input resistance:

$$R_{in C} = (\beta + 1) R_i \quad (9.44)$$

Because $R_i \gg r_e$, the resistance $R_{in C}$ is much higher than $R_{in D}$.

9.6.5. Dynamic range. Under the dynamic range one understands a ratio of the maximum to the minimum input signal, expressed in decibels. A minimum signal is limited by intrinsic (internal) noise, and a maximum signal by nonlinear distortion of the signal waveform. A maximum permissible signal can approximately be estimated using the criteria of cutoff or saturation of a transistor.

Assume $V_c^0 = 1/2 E_c$ in the quiescent state. A positive input signal causes the potential V_c to drop down to zero (the transistor then

stays in saturation). With the negative polarity of an input signal, the potential V_c rises and approaches E_c (the transistor then goes off). So, in both cases the maximum increment ΔV_c reaches $1/2 E_c$. Dividing this value by the voltage gain of Eq. (9.38) gives the maximum permissible input signal:

$$V_{in \max} = \varphi_T \quad (9.45)$$

The signals equal to φ_T are practically unacceptable, because at the edges of the range the emitter current changes heavily, along with the resistance r_e and the gain. This results in large nonlinear distortion. *For the signal distortion to be small, signal amplitudes must lie within $0.5\varphi_T$.*

Common-mode signals can be far larger than differential signals since K_{CC} is much smaller than K_{DD} . Let us write the relation between the collector potential and common-mode signal, taking V_c^0 equal to $1/2 E_c$:

$$V_c = 1/2 E_c + K_{CC} V_{in c}$$

Substituting $V_c = V_{in c}$ (saturation condition) or $V_c = E_c$ (cutoff condition) into the right side, we obtain the maximum positive and negative common-mode signals respectively:

$$V_{in c \max}^+ = E_c / (1 - K_{CC}) \quad (9.46a)$$

$$V_{in c \max}^- = E_c / 2K_{CC} \quad (9.46b)$$

where $K_{CC} < 0$ [see Eq. (9.39)]. Because $|K_{CC}|$ is usually smaller than unity, it is easy to see that common-mode input signals can reach a few volts and even approach the values close to E_c .

9.6.6. Offset and drift compensation. In a practical differential amplifier, parasitic voltages and currents, which are present in the quiescent state, affect the output signal in the process of amplifying an ac component.

Inevitable asymmetry of the circuit branches in a practical *DA* is the cause of a finite output voltage difference $V_{c1}^0 - V_{c2}^0$ (**unbalance**) in the quiescent state. An equivalent differential signal at the input that corresponds to the above output voltage

$$V_{off} = (V_{c1}^0 - V_{c2}^0) / K \quad (9.47)$$

is called the *input offset voltage*. To eliminate the unbalance of output potentials and thus bring the output voltage to zero, one must apply an input differential signal equal to V_{off} but of the opposite sign.

The offset voltage consists of a few components, each being dependent on the spread in such quantities as emitter currents I_e , collector resistance R_c , and others.

The spread in emitter currents (with voltages V_e being equal) results from the spread of thermal currents in emitter junctions [see Eq. (4.36b)]: the smaller the current I_{e0} , the smaller will be the current I_e . To balance out the emitter currents, an "equalizing" differential signal V_{off1} should be fed to the input. This signal must have such a polarity and such a value that the voltage V_e in a transistor with smaller current I_{e0} should rise in magnitude, while V_e in a transistor with higher current I_{e0} should decrease to make the emitter currents equal:

$$\begin{aligned} V_{e1} &= V_e + 1/2 V_{off1} = \varphi_T \ln (I_e/I_{e01}) \\ V_{e2} &= V_e - 1/2 V_{off1} = \varphi_T \ln (I_e/I_{e02}) \end{aligned}$$

Subtracting the second equality from the first gives

$$V_{off1} = \varphi_T \ln (I_{e02}/I_{e01}) \quad (9.48)$$

Thus, if thermal currents differ by 20%, then $V_{off1} \approx 5$ mV.

The next important component of offset voltage is attributed to the spread in R_c . Let currents in both branches be the same. The difference between collector potentials in the quiescent state will then be equal to

$$V_{c1}^0 - V_{c2}^0 = \alpha I_0 \Delta R_c$$

Dividing this potential difference by the gain (9.37b) and substituting $r_e = \varphi_T/(1/2I_0)$, we find the second component of offset voltage:

$$V_{off2} = 2\varphi_T (\Delta R_c/R_c) \quad (9.49)$$

For example, if $\Delta R_c/R_c = 0.02$, then $V_{off2} \approx 1$ mV. A smaller value of this component over the first [compare with the example to Eq. (9.48)] is typical of differential amplifiers. Other components associated with the spread in α , r_c , etc., are still less significant.

It should be kept in mind that the offset voltage is temperature dependent. This dependence is characterized by temperature sensitivity, or temperature drift ε_V , which is commonly expressed in $\mu\text{V}^\circ\text{C}^{-1}$.

For the main component of offset voltage V_{off1} (resulting from the unbalance of emitter currents) the temperature drift may be estimated as the difference between the temperature drifts in V_{e1} and V_{e2} with the emitter currents balanced out [see formulas preceding Eq. (9.48)]. Using Eq. (3.23), it is easy to find

$$\varepsilon_V = \varepsilon_1 - \varepsilon_2 = V_{off1}/T \quad (9.50)$$

For example, if we set $V_{off1} = 5$ mV and $T = 300$ K, then $\varepsilon_V \approx 17 \mu\text{V}^\circ\text{C}^{-1}$.

From Eq. (9.50) it is obvious that the temperature drift ε_V decreases with a decrease in offset voltage. However, at $V_{off1} < 1$ mV

such a proportionality is upset because the main component of offset voltage becomes the quantity V_{off2} arising from the unbalance of resistances R_c [see Eq. (9.49)]. The temperature drift in V_{off1} is the function of TC R_c , that is, it depends on the type and structure of resistors.

Along with the initial unbalance of collector potentials, there also exists an initial unbalance of input (base) currents ΔI_{in} . This parameter is known as an *input-offset current*, or just the *input current difference*.

In an ideal DA, the currents in both branches are equal, $\Delta I_{in} = 0$. In a practical DA, there is a difference between these currents that may be written in the form

$$\Delta I_{in} = \frac{I_{e1}}{B_1 + 1} - \frac{I_{e2}}{B_2 + 1} \approx \frac{I_{e1}}{B_1} - \frac{I_{e2}}{B_2}$$

Using relations $I_{e1} > I_{e2}$ and $B_1 < B_2$ specific to most unfavourable conditions we can readily obtain:

$$\Delta I_{in} = \frac{I_0}{B} \left(\frac{\Delta I_e}{I_e} + \frac{\Delta B}{B} \right) \quad (9.51)$$

where I_e and B are mean values ($I_e = 1/2 I_0$). As is apparent, the offset current decreases with a decrease in the working current of a DA and with an increase in the gain B .

A significance of parameter ΔI_{in} lies in that the offset current flowing through a differential-signal source resistance R_{gD} produces a voltage drop $\Delta I_{in} R_{gD}$. This voltage is equivalent to the offset voltage, or false signal. If $\Delta I_{in} = 20$ nA and $R_{gD} = 100$ k Ω , then $\Delta I_{in} R_{gD} = 2$ mV.

One more parameter of a DA that plays an important part is an *average input current*:

$$I_{in\ av} = 1/2 (I_{b1} + I_{b2})$$

In estimating this parameter, we may use the mean values of base currents, $I_{b1} = I_{b2} \approx I_e/B$. So,

$$I_{in\ av} = I_0/2B \quad (9.52)$$

The approaches to decreasing $I_{in\ av}$ are the same as for ΔI_{in} .

From the comparison of Eqs. (9.51) and (9.52) it follows that the offset current is much smaller than the average input current, typically by a factor of 10.

The significance of $I_{in\ av}$ lies in that this current flowing through a common-mode signal source resistance R_{gC} produces a voltage drop equivalent to the common-mode signal. Being multiplied by K_{DC} , this signal causes an initial unbalance of potentials at the output.

Both the average input current and the input offset current depend on temperature, that is, each exhibits its own temperature drift.

From Eqs. (9.51) and (9.52) it is clear that these drifts are dependent first of all on the B-T relationship. The temperature drifts of current parameters are proportional to the parameters proper, so that a decrease in the latter entails an increase in temperature stability.

9.6.7. Transients. The character of transients in a DA is the same as in a simple amplifying stage (see Fig. 9.7). The analysis performed in Sec. 9.4 keeps valid here too. But the quantitative parameters such as the time constant and rise time prove worse than in a simple stage.

Write the time constant τ_{hf} in the same form as that of Eq. (9.16):

$$\tau_{hf} = \tau_{oe}/(1 + \beta\gamma_b)$$

In Eq. (9.17) for γ_b we set $R_e = 0$. Whence,

$$\gamma_b = r_e/(r_e + R_g + r_b)$$

Other things being equal, the factor γ_b in a DA is apparently smaller than for a simple amplifier, and hence the time constant τ_{hf} is larger. Thus if $r_e = 25 \Omega$, $r_b = 150 \Omega$, $\beta = 100$, $R_g = 1 \text{ k}\Omega$, and if $R_e = 2 \text{ k}\Omega$ in the simplest stage, then the speed of response of a DA proves a factor of 20 lower than for the simplest stage.

With a decrease in the working current, the resistance r_e grows in accordance with (4.41) and so does the factor γ_b . In the limit, when $\gamma_b \approx 1$, the time constant τ_{hf} will be $1/(\beta + 1)$ as large as τ_{oe} ; in other words, it converts to the equivalent time constant $\tau_{\alpha oe}$ given by Eq. (9.11). But this does not imply that a substantial change in the speed of response takes place, because a decrease in current, as follows from Eq. (9.2b), must involve an increase in R_c , and the latter quantity determines the value of $\tau_{\alpha oe}$.

A practical approach to increasing the speed of response is that defined in Sec. 9.4, which presupposes a decrease in signal-source resistance and improvement in hf parameters of a transistor.

9.6.8. Voltage multiplier mode. The task of multiplication is to obtain an output voltage proportional to the product of two input voltages:

$$V_{out} = kV_{in1}V_{in2} \quad (9.53)$$

where k is the proportionality factor.

There are few circuit versions of multipliers. One of the popular types uses a differential amplifier. Replacing the quantity r_e in Eq. (9.37b), by φ_T/I_e according to Eq. (4.41) and taking into account the relation $I_e = I_0/2$, we may write the output voltage in the form

$$V_{out} = KV_{in1} = -\frac{\alpha R_c}{2\varphi_T} I_0 V_{in1} \quad (9.54)$$

where the subscript "1" identifies the input voltage. As seen, it is sufficient to ensure the dependence $I_0 \sim V_{in2}$ for the DA to enable it to perform the function of a multiplier. In other words, it is necessary to control the current I_0 which so far has been considered invariable.

A typical current source for the DA is a transistor connected in the circuit as shown in Fig. 9.35a. Here the element L is a load; in the given case, this is a DA emitter circuit (the lower terminal of L corresponds to the point E in Fig. 9.11). The element D is a silicon (reference) diode providing a base bias E_0 and determining the value of I_2 (I_0 in the given case). If we now apply a sufficiently large

signal $V_{in2} \gg V^*$ instead of the bias E_0 , the expression for I_0 will assume the form

$$I_0 = \frac{V_{in2} - V^*}{R_0} \approx \frac{V_{in2}}{R_0}$$

Substituting this expression into (9.54) gives the relation (9.53), in which

$$k = -\frac{\alpha R_c}{2\varphi_T R_0}$$

Thus, the DA with a controlled gain¹ is in principle capable of voltage multiplication. But the parameters of such a simple multiplier turn out to be not high enough. First of all, the device shows a significantly limited dynamic range: the relation (4.41) is only valid if $V_{in1} < \varphi_T$. The voltage V_{in2} is a common-mode signal which, as known, "leaks" through to the output and distorts small ac signals. The polarity of V_{in2} is also limited since it must be positive.

By elaborating a DA circuit, it becomes possible to obviate the above limitations to a considerable degree. Such circuits perform multiplication functions over a large dynamic range to a high accuracy. Note that multipliers can be used not only to perform a mathematical operation (9.53), but also amplitude-modulated hf voltage by lf voltage. This is one of the most important problems involved in communication and radio engineering.

By elaborating a DA circuit, it becomes possible to obviate the above limitations to a considerable degree. Such circuits perform multiplication functions over a large dynamic range to a high accuracy. Note that multipliers can be used not only to perform a mathematical operation (9.53), but also amplitude-modulated hf voltage by lf voltage. This is one of the most important problems involved in communication and radio engineering.

9.6.9. MOS transistor differential amplifier. The circuit of a simple DA based on MOS transistors appears in Fig. 9.13. The branches of this DA are simple dynamic-load stages (see Fig. 9.8b).

Since the differential signal V_{in} is equally divided between the gate-drain regions of active transistors $T1$ and $T3$, the voltage gain

¹ The controlled gain here is often referred to as "controlled transconductance" because for a transistor, $S = \alpha/r_e$ (see p. 161).

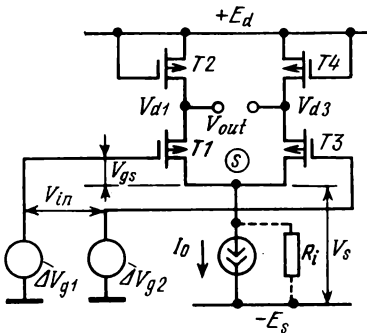


Fig. 9.13. MOS differential amplifier

of the DA may be considered to be the same as for the individual stage:

$$K = -S_1/S_2 = \sqrt{B} \quad (9.55)$$

where the parameter $B = b_1/b_2$ characterizes the geometry of transistors $T1$ and $T2$ [see Eq. (9.27)]. As with a simple stage, this DA has a limited value of K , typically 5 to 7, which is much lower than for bipolar DAs.

Determine the common-mode gain using the following relationships. A common-mode signal causes common-gate and gate-drain voltage increments (see Fig. 9.13):

$$V_{in\ C} = \Delta V_{gs} + \Delta V_s$$

The voltage ΔV_{gs} gives the increment

$$\Delta I_d = S \Delta V_{gs}$$

and the voltage ΔV_s the increment

$$\Delta I_0 = \frac{\Delta V_s}{R_i} = \frac{V_{in\ C} - \Delta V_{gs}}{R_i}$$

The increment ΔI_0 is distributed equally between the two branches of the DA, and so $\Delta I_d = 1/2 \Delta I_0$. Substituting the above expressions for currents into this equality, we find

$$\Delta V_{gs} = V_{in\ C} / (2SR_i + 1) \quad (9.56)$$

Multiplying ΔV_{gs} by K and dividing by $V_{in\ C}$ produces the gain for the common-mode component:

$$K_{CC} = K / (2SR_i + 1) \quad (9.57)$$

The gain K_{CC} is obviously smaller than K (usually by a few orders of magnitude) and generally comes to merely fractions of unity.

The gain K_{DC} that forms the basis of the common-mode rejection ratio results from the following. Let the DA branches be asymmetric so that $K_1 \neq K_3$ (the subscripts refer to $T1$ and $T3$). In this case the differential output signal caused by the common-mode input signal can be written as

$$V_{out\ D} = (K_1 - K_3) \Delta V_{gs} = (K_1 - K_3) \frac{V_{in\ C}}{2SR_i + 1}$$

where ΔV_{gs} is such as given by Eq. (9.56).

Setting $K_1 = K + \Delta K$, $K_3 = K - \Delta K$ (where K is the mean value), and dividing $V_{out\ D}$ by $V_{in\ C}$, we obtain

$$K_{DC} = 2\Delta K / (2SR_i + 1) \quad (9.58)$$

Find the common-mode rejection ratio dividing K by K_{DC} .

Write it in the general form analogous to Eq. (9.42):

$$K_R = (2SR_i + 1)/2\delta \quad (9.59)$$

Here δ is the asymmetry coefficient, which in the given case has the form

$$\delta = \Delta K/K$$

Using Eq. (9.55), we can represent the coefficient δ as a sum of individual spreads explained above in describing Eq. (9.42).

From Eq. (9.59) it is clear that the basic way for raising K_R is to increase the transconductance of active transistors and the resistance of a current source. For example, if $S = 1 \text{ mA/V}$, $R_i = 0.1 \text{ M}\Omega$, and $\delta = 0.04$, then K_R will equal 2 500 (about 70 dB). This parameter of MOS transistor DAs is generally smaller than for bipolar DAs.

Input resistances (both for differential and for common-mode signals) may practically be thought of as being infinitely large; they usually range between 10^{10} and $10^{12} \Omega$. The input currents are respectively negligible. From this it follows that *such parameters of a DA as the input-offset current (the difference between input currents), mean input current, and their temperature drifts are not limiting factors when using MOS differential amplifiers.*

In bipolar DAs, the main component of voltage V_{off} stems from the spread in emitter thermal currents (see Subsec. 9.6.6). In MOS differential amplifiers, this component is due to the spread in threshold voltages and specific transconductances of active transistors, that is, the parameters dependent not only on the geometry and electrophysical properties of a chip, but also on the conditions of its surface [the specific transconductance depends on the surface conditions through the surface carrier mobility as clear from Eq. (5.7)]. As known, it is more difficult to control the surface conditions than the bulk properties of a chip and its geometry. The voltage V_{off} is therefore higher than for bipolar differential amplifiers.

The transient response of MOS differential amplifiers is the same as for simple stages that form the branches of a DA circuit. The requisite expressions for transients were given in Subsec. 9.5.3.

9.7. Emitter Followers

Followers are amplifiers with a voltage gain close to unity, which reproduce the input signal without polarity reversal and present an increased input resistance and decreased output resistance as against simple amplifying stages.

The classical circuit of an *emitter follower* (common-collector amplifier) and its small-signal model appear in Fig. 9.14a and b respectively. It is easy to see that the emitter follower differs from

the simplest amplifier of Fig. 9.3a only in that its output voltage is taken off the emitter rather than off the collector and that its collector circuit has no resistor R_c .

9.7.1. Voltage gain. Inspection of the circuit in Fig. 9.14a indicates that if $R_g = 0$ and $V^* = \text{constant}$, then the input signal fully

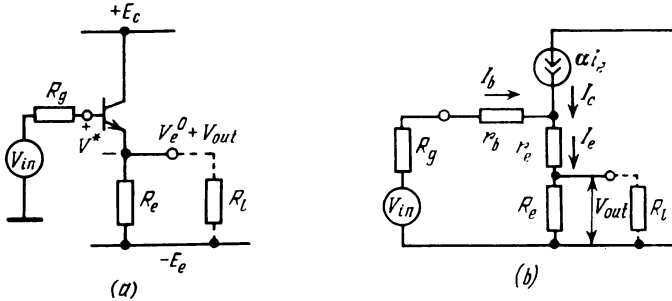


Fig. 9.14. Schematic diagram (a) and small-signal circuit model (b) of an emitter follower

goes to the output: $V_{out} = V_{in}$, and so $K = 1$. Considering the resistances R_g , r_b , and r_e , from Fig. 9.14b we obtain the relation

$$V_{out} = V_{in} - I_b (r_b + R_g) - I_e r_e$$

where $V_{out} = I_e R_e$ and $I_b = (1 - \alpha) I_e$.

From this relation it is easy to find the value of I_e and then express V_{out} in terms of V_{in} . The voltage gain of a follower will then assume the general form

$$K = \frac{R_e}{R_e + r_e + (1 - \alpha)(r_b + R_g)} \quad (9.60)$$

For example, if $R_g = 0$, $R_e = 5 \text{ k}\Omega$, $r_e = 25 \text{ }\Omega$, $r_b = 150 \text{ }\Omega$, and $\beta = 100$, then K will be near 0.995. If $R_g = 2 \text{ k}\Omega$, then K decreases a little to about 0.991.

With an external load resistance R_L connected to the circuit, as shown by a dash line in Fig. 9.14, the gain drops off still more because of the replacement of R_e by $R_e \parallel R_L$.

From Eq. (9.60) it is seen that the voltage gain is positive. This means that *the follower does not reverse the polarity of an input signal* or, in the case of a sinusoidal signal, *does not reverse its phase* (at sufficiently low frequencies, of course).

Despite the fact that the follower never provides a voltage gain in excess of unity, the device belongs to the class of amplifiers because it *amplifies current*. This is obvious from the well known relation between the output (emitter) current and input (base) current: $I_e = (\beta + 1) I_b$, where $\beta \gg 1$.

9.7.2. Input resistance. The input resistance of the follower shown in Fig. 9.14 is described by the same formulas (9.7) as for the simplest amplifier. In the general case,

$$R_{in} = (\beta + 1) (R_e + r_e) + r_b \quad (9.61a)$$

and in a particular case, where r_e and r_b may be disregarded,

$$R_{in} = (\beta + 1) R_e \quad (9.61b)$$

The external load resistance R_l being allowed for, the resistance R_{in} becomes smaller.

In distinction to an amplifying stage, the follower permits an increase in R_e , along with the input resistance, without practically changing the voltage gain [cf. Eqs. (9.5a) and (9.60)]. A rise in R_e ,

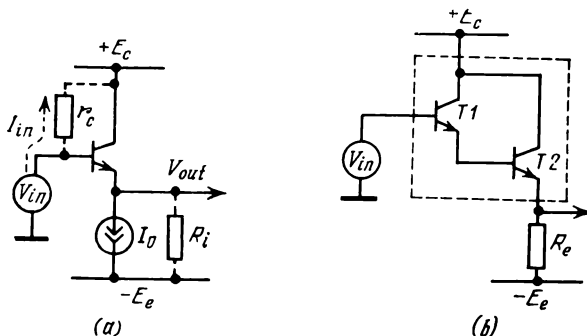


Fig. 9.15. Emitter followers with increased input resistance
(a) having current source as load; (b) using Darlington pair

however, necessitates an increase in supply voltage E_e to retain the desired value of quiescent current I_e^0 given by Eq. (9.2a). The limitations of such an approach to increasing the input resistance are obvious. A practical approach is to insert a current source into the emitter circuit or to use a composite transistor.

Figure 9.15a illustrates an emitter follower with a current source I_0 in place of R_e . If the current source is ideal ($R_i = \infty$), then $R_{in} = \infty$ according to Eq. (9.61). In reality, the input resistance has a finite value conditioned by the collector junction resistance (shown by a dash line in Fig. 9.15a). So far this resistance has been disregarded because its role is insignificant in conventional amplifiers. But in the given case, where the emitter circuit is "open" for ac components (because the emitter current has a fixed value), the resistance r_c represents the only element through which the input current can flow. So the maximum value of input resistance for a follower (and any other amplifier)

$$R_{in \max} = r_c \quad (9.62)$$

At $I_e = 1$ mA, r_c comes to 2 or 3 M Ω . As the current decreases, r_c rises according to Eq. (4.42), but its upper limit depends on leakage currents through the collector junction.

At a finite current-source resistance R_i , the input resistance of a follower will be lower than r_c . The input resistance can be regarded as r_c connected in parallel to $(\beta + 1) R_i$.

Figure 9.15b shows the follower using a Darlington transistor (see Fig. 9.1a). A composite transistor is known to feature a rather high current gain $\beta \approx \beta_1 \beta_2$ [see Eqs. (9.1)]. According to Eqs. (9.61), therefore, in the follower connected in a Darlington circuit, *it is easy to obtain a high input resistance at a comparatively small resistance R_e* . For example, if $R_e = 2$ k Ω and $\beta = 2$ 000, then the **calculated** value of R_{in} would be equal to about 4 M Ω . As with the preceding circuit, this circuit has the real value of R_{in} limited by r_c [see Eq. (9.62)].

9.7.3. Output resistance. This resistance can be found proceeding from the general definition given on p. 319. First, set $R_i = \infty$ (see Fig. 9.14). In this case,

$$(V_{out})_{oc} = K V_{in}$$

where K is the voltage gain of Eq. (9.60). Set now $R_i = 0$, that is, short out the resistance R_e . Then,

$$(I_{out})_{sc} = \frac{V_{in}}{r_e + (1 - \alpha)(R_g + r_b)}$$

[see, for example, Eq. (9.4)]. Dividing the open-circuit voltage by the short-circuit current and substituting the expression for K , we can readily present the output resistance in the form

$$R_{out} = R_e \parallel [r_e + (1 - \alpha)(R_g + r_b)] \quad (9.63a)$$

The external component R_e is usually of little significance as also is the component $(1 - \alpha)r_b$ against r_e . For practical emitter follower circuits, therefore, we may use a simplified expression

$$R_{out} = r_e + \frac{1}{\beta + 1} R_g \quad (9.63b)$$

As seen, in the general case the output resistance depends on the signal source resistance. But at sufficiently large values of β (when using the Darlington pair, for example) the addend in Eq. (9.63b) may be omitted. The output resistance is then **minimum**, being determined only by the emitter junction resistance:

$$R_{out \min} = r_e \quad (9.64)$$

It can be readily noticed that the ratio of the input to the output resistance in a follower is significantly larger than for a simple amplifying stage and differential amplifier, in which this ratio does not

exceed the value $\beta + 1$. The ratio between the **maximum** input resistance of Eq. (9.62) and the minimum output resistance of Eq. (9.64) is r_c/r_e , typically over 50 000. This ratio is independent of working current since both r_c and r_e are in inverse proportion to current.

Because of the large difference between the input and output resistances, the follower finds wide use as a *buffer stage*, or an *impedance transformer*. This function is illustrated in Fig. 9.16.

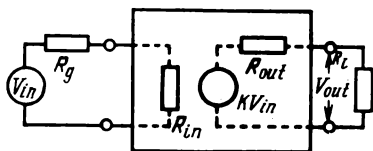


Fig. 9.16. Emitter follower as a buffer stage

Assume there is a signal source V_{in} with resistance R_g and a load with resistance R_l , where $R_g \gg R_l$. Connecting the load **directly** to the source V_{in} produces a very low voltage gain:

$$\frac{V_{out}}{V_{in}} = \frac{R_l}{R_g + R_l} \ll 1$$

If we now connect the buffer stage of Fig. 9.16 between the source V_{in} and load, making sure that the relations $R_{in} \gg R_g$ and $R_{out} \ll \ll R_l$ typical of the emitter follower hold good, the circuit will maintain a voltage gain of near unity. So, the buffer stage can artificially, as it were, decrease the signal source resistance or increase the load resistance. Hence, it really acts as an impedance transformer (impedance-matching circuit).

In particular, a buffer stage can be used to reduce R_g at the input of a differential amplifier and thus raise its speed of response (see Subsec. 9.6.7). Buffer stages also find extensive use where capacitive loads are present (cables, long wires, intricate interconnection patterns in ICs). At rather high frequencies, the impedance of a capacitive load becomes small, and its direct connection to a high-resistance signal source results in a decreased voltage gain, as mentioned earlier. A buffer stage enables improving the transmission of a signal at high frequencies and widening the operating frequency range of amplifiers.

9.7.4. Transients. The equivalent circuit for the analysis of transients is presented in Fig. 9.17. The formal analysis of this circuit is simple, but leads to awkward and poorly illustrative expressions. Therefore, it is more useful to consider characteristic features of a transient process and apply approximate calculation formulas. Let us initially disregard capacitances C_c and C_l and apply a step signal

V_{in} to the input. Since at the first moment the collector current does not vary (the current generator stays inactive), the circuit converts to a passive resistive network. This means that some time after application of the signal the circuit does not provide the gain in power.

The initial currents and output voltage are given by

$$I_{in}(0) = I_b(0) = I_e(0) = \frac{V_{in}}{R_g + r_b + r_e + R_e} \quad (9.65a)$$

$$V_{out}(0) = I_e(0) R_e = V_{in} \frac{R_e}{R_g + r_b + r_e + R_e} \quad (9.65b)$$

The initial voltage gain is

$$K(0) = R_e / (R_g + r_e + r_b + R_g) \quad (9.66)$$

As a certain time elapses, a step $I_e(0)$ causes the collector current to augment. The increments in I_c are distributed between the emitter and base circuits in inverse proportion to their resistances.

If the signal source resistance is sufficiently small ($R_g \ll R_e$), the increments ΔI_c mainly go to the base circuit. Correspondingly, the

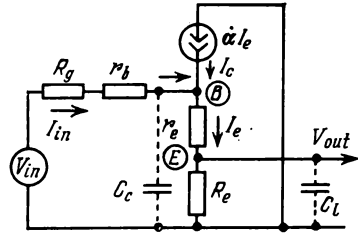


Fig. 9.17. High-frequency emitter follower circuit model

emitter current and hence the output voltage change insignificantly and remain close to the initial values (see solid lines in Fig. 9.18a). The rise time t_{r1} taken between the 10% and 90% levels of the steady-state current proves equal to zero; in other words, the leading edge builds up instantly. But the transients in the collector and the base circuit proceed in a rather clearly defined manner: the collector current grows from zero to the steady-state value αI_e and the base current drops from the initial value I_e to $(1 - \alpha) I_e$. The ratio between I_e and I_b thus rises, that is, after a certain time the follower begins to amplify power. Since the emitter current remains almost constant, the collector and base currents both change with a time constant τ_α .

If the condition $R_g \ll R_e$ is not met, the transient proceeds in a somewhat different way (see dash curves in Fig. 9.18a); namely, the initial surges of the emitter current and output voltage become smaller and the subsequent increments of the collector current branch

out not only into the base circuit but also, to a large extent, into the emitter circuit. The current I_e and voltage V_{out} then noticeably grow during the transient. As seen from the figure, t_{r2} here has a finite value. The collector current and also other quantities change at a time constant given by Eq. (9.13). But in this expression it is necessary to replace $\tau_{\alpha o e}$ by τ_α , since at the beginning of the analysis we have neglected C_c . Thus,

$$\tau_{hf} = \tau_\alpha / (1 - \alpha\gamma_e) \quad (9.67)$$

Commonly, τ_{hf} ranges from 2 to $5\tau_\alpha$.

Allow for C_c now, but first set $\tau_\alpha = 0$, that is, disregard the time lag of the processes in the base. On applying a step input, the input current first fully flows through the capacitance, while the transistor currents and output voltage remain constant. As the capacitance charges, the transistor currents and output voltage smoothly grow to the steady-state values (see solid curves in Fig. 9.18b).

The time constant of the transient is determined by C_c and resistances shunting this capacitance (see Fig. 9.17):

$$\tau_c = C_c [(R_g + r_b) \parallel (R_e + r_e)] \quad (9.68)$$

The leading edges of pulses will grow exponentially; the pulse rise time from the 10% to 90% level of the steady-state value is

$$t_r = 2.2\tau_c \quad (9.69)$$

The time constant τ_c is typically equal to about 0.5 to 1 ns. With an ideal signal source ($R_g = 0$), τ_c practically coincides with the base time constant given by Eq. (4.65): $\tau_c = C_c r_b$. The values of this quantity range from 0.1 to 0.2 ns and below.

The initial "peak" of input current (see Fig. 9.18b) by far exceeds the steady-state value of base current. Indeed, the voltage across C_c does not change at the first moment, and hence the initial input current is dependent on the input voltage and base circuit resistance:

$$I_{in}(0) = V_{in} / (R_g + r_b)$$

For example, if $V_{in} = 0.1$ V and $R_g + r_b = 1$ k Ω , then $I_{in}(0) = 0.1$ mA; in the steady state, however, at $R_{in} = 100$ k Ω , the current is $I_{in} = 1$ μ A, or one-hundredth as large.

Show now that formula (9.69) holds for the approximate estimation of a pulse rise time with regard to **both** inertial factors such as the collector capacitance and the transit time for carriers in the base. For this, consider two limiting cases: $\tau_c > 2\tau_{hf}$ and $\tau_c < 0.5\tau_{hf}$. The first is specific to comparatively high-resistance signal sources, and the second to comparatively low-resistance sources.

For the first case, τ_{hf} is of little significance. So the transient is characterized by τ_c and the rise time is defined by formula (9.69).

For the second case, the transient proves “composite”. Really, the currents I_b and I_e grow with a time constant τ_c , that is, faster than does the current I_c which rises with a time constant τ_{hf} . So it can be assumed that at the first stage the currents I_b and I_e both reach a maximum described by Eq. (9.65a), and at the second stage the current I_c , branching out into the low-resistance base circuit, reduces

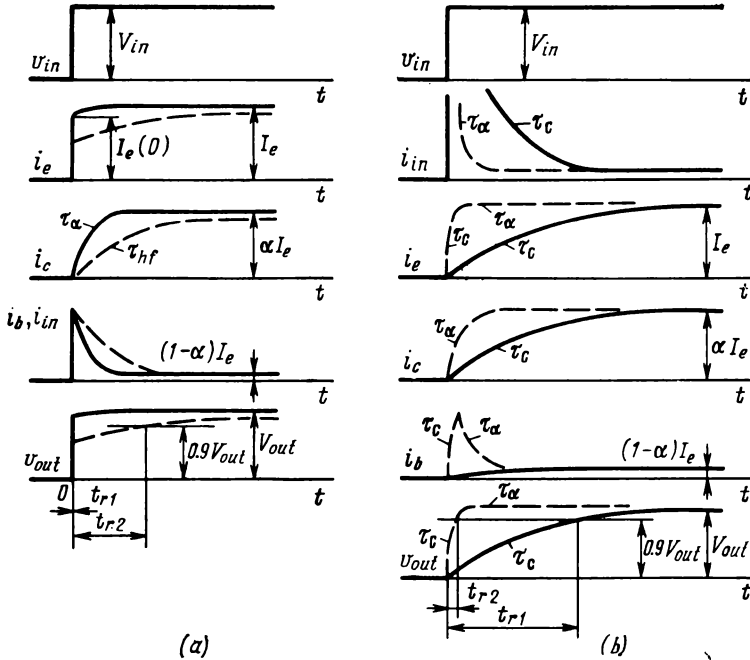


Fig. 9.18. Transients in an emitter follower with the collector capacitance disregarded (a) and considered (b)

the base current to the steady-state value. The emitter current and hence the output voltage remain almost constant at the first stage. Consequently, the rise time is determined by the *duration of the first stage*, that is, by formula (9.69).

For an intermediate case ($\tau_c \approx \tau_{hf}$), the transient is found to be more complicated, but the calculations show that formula (9.69) gives acceptable results here too.

If the load capacitance is rather small, namely, if $C_l R_e < \tau_c$, then it may be regarded as being connected in parallel with the collector capacitance. The rise time and the fall time then grow accordingly. But if the load capacitance is large, the rise time and the fall time may become different in value (see next subsection).

9.7.5. Lockout of a follower. A specific transient process takes place in a follower at a sufficiently large load capacitance C_L (see Fig. 9.17). At the first moment after arrival of a step input, the voltage across the capacitance C_L does not change, and hence the emitter potential remains constant. So, if the signal V_{in} is negative in sign and in excess of V^* (Fig. 9.19), the voltage at the emitter junction becomes reverse, which drives the transistor into cutoff.

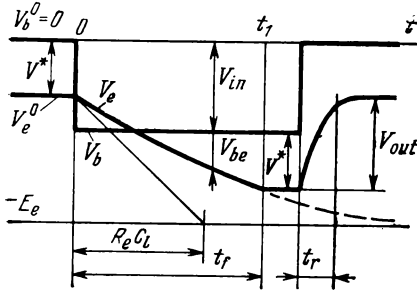


Fig. 9.19. Transients in an emitter follower at high load capacitance

Next, the capacitance C_L discharges via R_e with a large time constant $C_L R_e$. The transistor turns fully on only when the base-emitter voltage reaches V^* (at t_1 in Fig. 9.19). The described effect that causes the transistor to go temporarily off at large negative signals received the name of *follower lockout*.

As known, to render a silicon transistor nonconductive, the reverse bias on the emitter junction

is not obligatory: it is enough to have the forward bias 0.1 or 0.2 V below the value V^* . Therefore, lockout can occur even at small negative signals, 0.2 or 0.3 V.

As a result of lockout, the fall time exceeds the rise time, the latter being determined by formula (9.69) as before. As regards the fall time, this can be readily found from the expression describing the discharge of a load capacitance via R_e (see Fig. 9.19):

$$v_e(t) = -V^* e^{-t/\tau_l} - E_e (1 - e^{-t/\tau_l})$$

where $\tau_l = C_L R_e$. Substituting $V_e = -V_{in} - V^*$ (the on condition for the emitter junction at t_1 in Fig. 9.19) into the left side of the expression, we find

$$t_f = \tau_l \ln \left(1 - \frac{V_{in}}{E_e - V^*} \right)^{-1} \quad (9.70)$$

For example, if $\tau_l = 40$ ns, $E_e = 2$ V and $V_{in} = 0.5$ V, then t_f would equal about 24 ns. This value of t_f is much higher than the typical values of t_r .

It should be pointed out that the follower lockout and the difference in t_r and t_f presuppose the application of a step input; otherwise C_L has time to "follow up" a negative input signal and the lockout does not occur. A criterion for lockout may be an inequality

$$t_{f\text{ in}} < r_b C_L$$

where $t_{f\text{ in}}$ is the input pulse fall time.

9.7.6. Level shifters. In a multistage amplifier, the base of each successive stage receives not only a useful signal but also a dc voltage component from the collector of the preceding stage. The dc component "accumulates", grows from stage to stage, which causes difficulties in evolving the last, output stages. A problem thus often emerges which calls for elimination of a dc component at the input of the next stage without changing, where possible, the ac component (signal). It is *level shifting circuits* that solve this problem.

The simplest dc level shifter is an emitter follower. Really, it has the level of an output (emitter) potential that is lower than the base potential level by V^* , but amplifies the ac signal at $K \approx 1$.

The emitter follower provides the basis for other, more complex level shifters. For example, if there is a need to lower the level of an input signal by $2V^*$, it is possible to use either a Darlington pair (see Fig. 9.15b) or insert a forward-biased diode (see below) into the emitter circuit of the simplest follower.

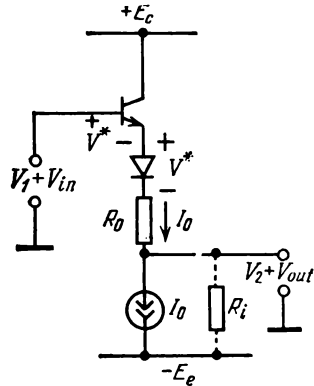


Fig. 9.20. Level shifter circuit

It is sometimes required to shift the level by the value that is not a multiple of V^* , for example, by 2.5 V. A universal level shifting circuit shown in Fig. 9.20 may serve the purpose. In the general case, such a circuit may include not one but n series-connected diodes. The relation between the input and output levels has the form

$$V_1 - V_2 = (n + 1) V^* + I_0 R_0 \quad (9.71)$$

Varying the values of n , I_0 , and R_0 can provide for any desirable shift of the level. Thus, if the desired shift is $V_1 - V_2 = 2.5$ V, the number of diodes to be chosen is $n = 2$. In this case, $(n + 1) V^* = 2.1$ V and $I_0 R_0 = 0.4$ V; at $I_0 = 1$ mA, the required value of the resistance is $R_0 = 400 \Omega$.

The ac voltage gain in the circuit of Fig. 9.20 primarily depends on the current-source internal resistance. If $R_i = \infty$, then $K = 1$ irrespective of the structure of the emitter and base circuits. But if R_i has a finite value, it should be substituted for R_e in (9.60). Besides, r_e should be replaced by $(n + 1) r_e + R_0$. Then,

$$K = \frac{R_i}{R_i + (n + 1) r_e + R_0 + (1 - \alpha)(r_b + R_g)} \quad (9.72)$$

The resistance R_i generally lies in the range of 100 k Ω and above, and other resistances in the denominator of Eq. (9.72) do not exceed 1 or 2 k Ω . For this reason, the ac voltage gain comes to near unity. Given the load resistance R_L , the value R_i in (9.72) should be replaced by a smaller value of $R_i \parallel R_L$. The circuit gain will then decrease accordingly.

9.8. Cascode Amplifier

The cascode is an amplifier circuit having two transistors connected in series so that the same current flows through each transistor (Fig. 9.21). This circuit configuration can be regarded as a unit (see the dash line), that is, as one of the versions of a composite transistor. The current gain α of this composite transistor is easy to find from the following obvious relationships:

$$I_{c2} = \alpha_2 I_{e2} = \alpha_2 I_{c1} = \alpha_2 (\alpha_1 I_{e1})$$

Dividing I_{c2} by I_{e1} yields

$$\alpha = \alpha_1 \alpha_2$$

The cascode configuration thus provides an emitter current gain that differs but little from that obtained in one (active) transistor $T1$. It is obvious, therefore, that the voltage gain of the cascode circuit will practically be the same as that for a simple amplifying stage [see Eq. (9.5b)]:

$$K \approx -\alpha_1 \alpha_2 (R_c/R_e)$$

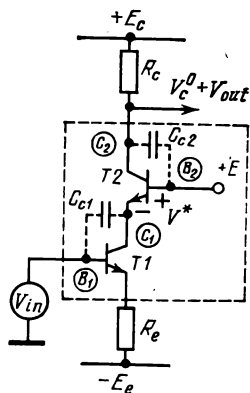


Fig. 9.21. Cascode

The cascode thus does not afford any improvement in the gain (and also in the input and output resistances), but offers an important advantage over a simple amplifier; namely, *it has no connection between the output point (collector C_2) and input point (base B_1)* and so provides the output-to-input isolation. In a simple amplifier, however, the output and input points are connected via C_c and r_c . Such a feedback path often complicates the operation of amplifiers. In particular, feedback causes an increase in the input capacitance of a stage (Miller effect, see Subsec. 9.5.4). In the presence of an inductive component in the load impedance, such a feedback path will lead to generation of parasitic oscillations, that is, conversion of the amplifier into an oscillator¹.

¹ The inductive component is inevitable where the load used is a tuned or tank circuit, as is the case with selective (tuned) amplifiers. Therefore, the problem of output-to-input isolation in tuned stages is particularly important.

The reason why the output stays *isolated* from the input is that the intermediate circuit point (base B_2) is at an invariable potential E . This potential may be regarded to be a supply voltage for $T1$. In this case, the load for this transistor is a rather small resistance r_{e2} of the emitter junction. So the transistor $T1$ practically operates with the collector circuit shorted out. Its gain is thus near zero, the Miller effect is nonexistent, and the input capacitance is equal to the transfer capacitance C_{c1} .

The output voltage of the cascode causes a flow of current through C_{c2} . But this current goes to "ground" via the fixed bias source E and does not get to the collector circuit of $T1$. For this reason, there is no feedback path between the output and input, so that the danger of parasitic oscillations is eliminated or, at least, greatly reduced. Owing to this feature, the cascode enjoys wide use in tuned amplifiers.

9.9. Output Stages

The objective of final stages is to deliver a specified (rather large) power and, hence, sufficiently large voltages and currents to the load. The voltage gain is a secondary parameter for output stages. The primary considerations are the *efficiency η and nonlinear distortion factor K_f* .

Output stages usually consume a major portion of the amplifier power; a high efficiency, therefore, implies an efficient use of the power source. This is of particular importance in integrated circuits where the power dissipated by a chip is limited. Nonlinear distortion is specific to the output stages for the following reason: a specified and sufficiently large power to be obtained at high efficiency inevitably requires the use of currents and voltages whose amplitudes are comparable to the values of dc components.

9.9.1. Parameters of output stages. The efficiency is defined as the ratio of ac output signal power over the dc power taken from the supply source:

$$\eta = \frac{1/2 V_{out\ m} I_{out\ m}}{E I_{av}} \quad (9.73)$$

where $I_{out\ m}$ and $V_{out\ m}$ are the amplitudes of output quantities.

Because the efficiency depends on output power, it is customary to define this ratio for a **maximum** ac power output over the **maximum** dc power input.

The distortion factor characterizes the difference in waveform between the output and the input signal. This difference results from a nonlinearity of the transfer characteristic of a stage (see Fig. 8.4). Nonlinear distortion shows up in that the output signal

contains new harmonics nonexistent in the input signal. The distortion analysis, therefore, customarily comes to calculating (or measuring) the harmonic content in the output for a purely sinusoidal input.

The energy characteristic of distortions is defined as the ratio of the total power of upper harmonics, starting from the second, to the power of the first (fundamental) harmonic whose frequency is equal to that of the input signal.

The distortion factor is usually defined as the square root of the ratio between the total power of upper harmonics at the output of an amplifier and the power of the first harmonic. If the load resistance is the same for all harmonics, then the powers are proportional to the currents or voltages squared. The distortion factor will then assume one of the two forms:

$$K_f = \frac{\sqrt{\sum I_m^2}}{I_1} = \frac{\sqrt{\sum V_m^2}}{V_1} \quad (9.74)$$

where m is the harmonic number, starting from the second.

The permissible value of K_f is dictated by the concrete requirements for the apparatus in question. In sound reproduction, for example, it is desirable that K_f be smaller than 2 or 3%. In measuring devices, K_f must be substantially smaller.

The analytical estimation of nonlinear distortion is possible only if the transfer characteristic is given as a mathematical function. The distortion factor is more often estimated graphically (using the known transfer characteristic) or experimentally (with special measuring devices).

9.9.2. Operation modes. Depending on the location of the initial operating point (quiescent point) on the transfer characteristic, one distinguishes a few classes of operation: A, B, AB, and others. These classes differ in the maximum values of efficiency and in the values of nonlinear distortion.

With *class A* operation, the operating point in the quiescent state lies in the center of the quasilinear portion of the transfer characteristic (Fig. 9.22). Obviously, the distortion here will be a minimum because both half-waves of the input signal are within the quasilinear region. The maximum efficiency can be found from formula (9.73) by substituting the maximum values of voltage and current amplitudes corresponding to the boundaries of the quasilinear region: $V_{out\ m} = 1/2 E_c$ and $I_{out\ m} = I^0 = I_{av}$. Thus,

$$\eta_m = 1/4 \text{ or } (25\%) \quad (9.75)$$

With the transistor transformer-coupled to the load, the maximum efficiency is doubled: $\eta_m = 50\%$. However, the use of transistors (as active components) in ICs is undesirable.

With *class B* operation, the quiescent point lies at a boundary of the quasilinear region, namely, at the boundary which corresponds to the off condition of a transistor (Fig. 9.23). It is clear that in

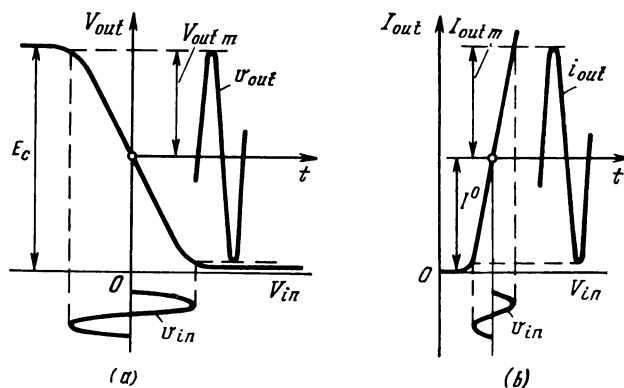


Fig. 9.22. Class A operation. Calculation of output voltage (a) and output current (b)

this operation, the output current flows only during the positive half-wave of the signal. The output voltage, therefore, proves rather nonsinusoidal, that is, contains a large number of upper harmonics.

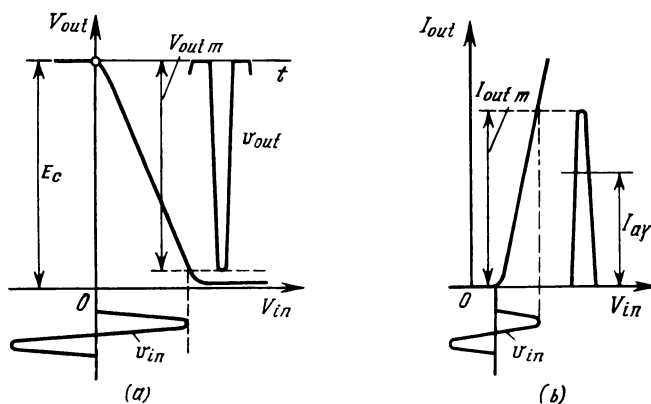


Fig. 9.23. Class B operation. Calculation of output voltage (a) and output current (b)

Analysis shows that *irrespective of the amplitude* of a signal, K_f reaches an impermissibly high value, 70%. Therefore, this operation mode in its simplest version is impracticable.

The class B operation is useful in a *push-pull circuit* consisting of two amplifiers, one of which amplifies the positive and the other

the negative half-wave of the signal. These half-waves delivered to the load add up to form a full sinusoid (Fig. 9.24). For the negative half-waves to be amplified, it is necessary either to connect to the input of amplifier 2 a phase-inverting element (a transformer, for example) or to use in the amplifier 2 a transistor of the other

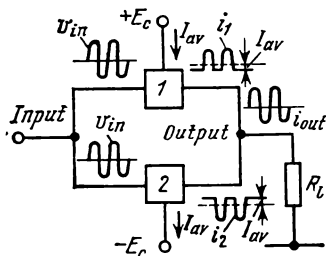


Fig. 9.24. Principle of a push-pull amplifier

type (*pnp* type). Examples of implementing a push-pull circuit are given in the next subsection.

The efficiency of a class B push-pull circuit can be estimated proceeding from the fact that the powers dissipated in each of its halves are equal. It is thus sufficient to calculate the power during one half-cycle. For this we should substitute the following quantities in Eq. (9.73):

$V_{out\ m} = E_c$ (see Fig. 9.23a), $E = E_c$, and $I_{av} = (2/\pi) I_{out\ m}$ (the

average value of sinusoidal current during the half-cycle, see Fig. 9.23b). The maximum efficiency will then take the following value:

$$\eta_m = \pi/4 \text{ or } (78\%) \quad (9.76)$$

For example, if an amplifier is fabricated on a chip capable of dissipating 300 mW, the useful power transferred to the load can be in excess of 1 W.

Class AB operation represents an intermediate case of class A and B operation: the initial operating point lies not at the cutoff boundary, but in the region of forward bias of the emitter junction, the forward currents being significantly lower than they are under class A conditions (an example will be given below).

9.9.3. Output stage circuits. In microelectronics, class A operation is unsuitable because it results in low efficiency. Most popular are push-pull amplifiers of class B and AB.

In Fig. 9.25a is shown a simple class B push-pull circuit based on **complementary** transistors. The load is connected to the emitter circuit of transistors, which thus act as voltage followers. Power gain is determined by the current gain.

In the quiescent state, both transistors stay cut off since the voltages at emitter junctions are equal to zero. During the positive half-wave of an input signal, the *nnp* transistor *T1* is on, and the current flows through the load as shown by a dash arrow 1. During the negative half-wave, the *pnp* transistor *T2* takes over and the current flows in the direction of a dash arrow 2. The output signal thus has two polarities as does the input signal. The power gain

is close to the ratio of the emitter to the base current, that is, near $B + 1$.

Unfortunately, this simple circuit is responsible for comparatively large nonlinear distortion. The cause of distortion is the presence of the "heel" on the input I - V characteristic for silicon transistors. The waveform of an output signal is easy to obtain graphically using the so-called *composite transfer characteristics* (Fig. 9.25b). The curve 1 refers to $T1$ and curve 2 to $T2$. As seen, the duration of the positive and the negative half-wave at the output is *shorter than the half-cycle*; these half-waves are separated by small horizontal

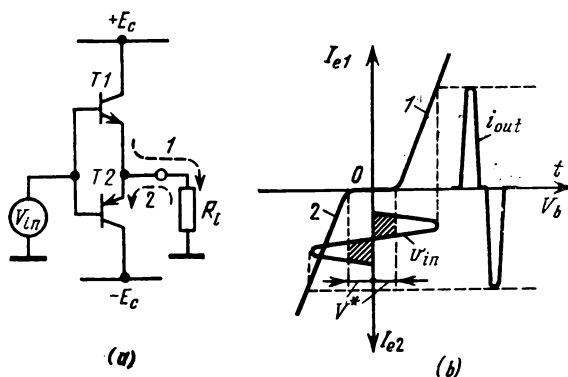


Fig. 9.25. Circuit (a) and transfer characteristics (b) of a class B push-pull amplifier using complementary transistors

portions (hatched areas here represent the parts of half-sinusoids left without amplification). Obviously, this type of distortion will be especially heavy at small input signals with an amplitude comparable to that of V^* .

To eliminate the above drawback, the amplifier circuit is made somewhat more complex by adding a level shifter (Fig. 9.26a) to supply a separate (individual) bias for each transistor base. The composite characteristics for this variant are given in Fig. 9.26b. The circuit operates under AB conditions.

Since the parameters of integrated *nnp* and *pnp* transistors differ heavily, it is common to use a composite *pnp* transistor (see Fig. 9.2) to serve the purpose of $T2$ shown in Figs. 9.25 and 9.26. This favours better symmetry of the output stage and reduces nonlinear distortion.

Should it be necessary to build the output stage around single-type (not complementary) transistors, the stage circuit will then appear such as shown in Fig. 9.27. The transistor $T1$ here is on throughout the input-signal cycle. In the quiescent state, I_{c1}^0 and R_c are set so that the collector potential V_{c1}^0 is equal to zero; the diode D and

transistor $T2$ stay off, and the current in the load does not flow.

During the positive half-wave of an input signal, the potential V_{c1} drops off, the diode D goes on, and the current flows through

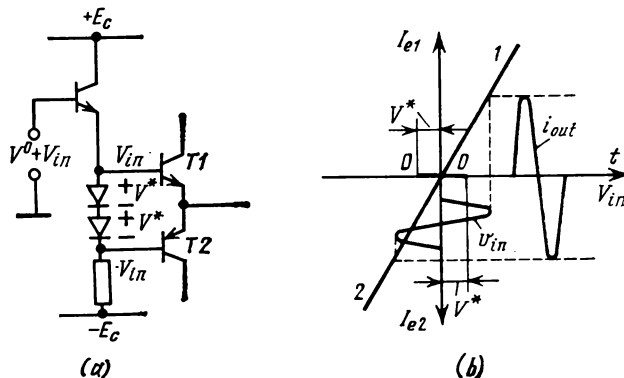


Fig. 9.26. Circuit (a) and transfer characteristics (b) of a class AB push-pull amplifier using complementary transistors.

the load as shown by a dash arrow 1. The transistor $T2$ stays off because the turn-on voltage V^* across the diode reverse-biases the emitter junction (see plus and minus signs unbracketed). During the negative half-wave, the potential V_{c1} rises, $T2$ switches on, and the current flows through the load as indicated by a dash arrow 2; the diode is off because the turn-on voltage V^* at the emitter junction reverses the diode into the reverse-biased condition (see plus and minus signs bracketed).

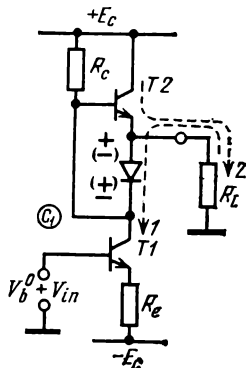


Fig. 9.27. Class AB push-pull amplifier using single-type transistors

ciently high, so there is commonly no need to secure class AB condition, that is, to apply an additional bias as was the case for circuit of Fig. 9.26.

The half-waves of voltage across the load will obviously be only if the gains of the positive and the negative signal are

$$V^*/K$$

where K is the gain of the stage using transistor $T1$. The stage gain can be

too. With a positive signal, the gain K^+ is proportional to R_l ; in this half-cycle, R_l is a collector load for $T1$. With a negative signal, K^- is proportional to R_c (in this half-cycle, $T2$ acts as an emitter follower). Consequently, for K^+ and K^- to be equal, the equality $R_c = R_l$ should be met.

A disadvantage of the circuit is that a change in R_l entails a change in the amplitude of a negative output voltage (when the load current flows in the direction of arrow I) because this entails a change in K^+ . To overcome this shortcoming, a buffer stage should be added to the circuit.

9.10. Voltage Regulators

Voltage regulators are primarily intended to act as supply sources for ICs. DC amplifiers are a good example: if they receive power from an unstabilized source, one of the drift components may grow too high (see Subsec. 9.3.3). Simple voltage regulators are also widely used as bias sources for many analog circuits, including operational amplifiers.

9.10.1. Voltage regulator parameters. The designations used in the skeleton diagram of Fig. 9.28 are as follows: V_1 is the unstabilized (input) and V_2 is the stabilized (output) voltage; I_1 and I_2 are the input and output currents; and $R_l = V_2/I_2$ is the load resistance.

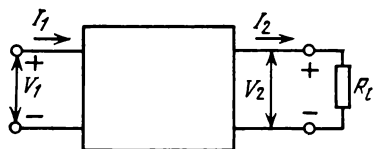


Fig. 9.28. Skeleton diagram of a voltage regulator

The output voltage of a regulator cannot of course be absolutely stable. The increments ΔV_2 are small (that is, lie within permissible limits), but yet depend on the increments in input voltage and output current¹. The relative instability of output voltage will then be expressed as a sum of two terms:

$$\frac{\Delta V_2}{V_2} = \frac{\Delta V_2^V}{V_2} + \frac{\Delta V_2^I}{V_2} \quad (9.77)$$

where the augend stems from the instability of input voltage, and the addend from that of output current. The components of relative

¹ Apart from these dependences, there also exists a temperature and time drift in output voltage that is analogous to the drift specific to dc amplifiers (see Subsec. 9.3.3). This type of instability will not be considered in the further discussion.

instability are customarily expressed in terms of two basic parameters of a regulator, known as the *stabilization factor* and *output resistance*¹:

$$\frac{\Delta V_2^V}{V_2} = \frac{1}{K_{st}} \frac{\Delta V_1}{V_1} \quad (9.78a)$$

$$\frac{\Delta V_2^I}{V_2} = -R_{out} \frac{\Delta I_2}{V_2} \quad (9.78b)$$

From Eq. (9.78a) it follows that the stabilization factor has the form

$$K_{st} = \frac{V_2}{V_1} \frac{\Delta V_1}{\Delta V_2} \quad (9.79)$$

(with the current I_2 considered constant). The general expression for output resistance (see p. 319) is

$$R_{out} = \frac{(V_{out})_{oc}}{(I_{out})_{shc}} \quad (9.80)$$

in using Eq. (9.80), it is necessary to represent a regulator by a small-signal circuit model. It is clear that the performance of a regulator improves as the stabilization factor increases and the output resistance decreases.

9.10.2. Diode regulators. A simple diode voltage regulator shown in Fig. 9.29a includes a current source and silicon reference diode. A small-signal model of the regulator is illustrated in Fig. 9.29b.

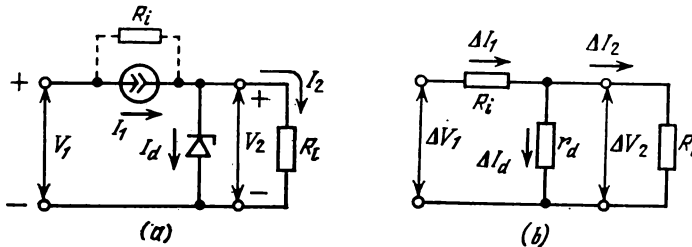


Fig. 9.29. Diode regulator with a reference diode

As is clear from the figure, the output voltage is determined by the rated voltage of the reference diode: $V_2 = V_d$. Commonly, $V_d > 5$ or 6 V in avalanche breakdown, and $V_d = 2$ to 5 V in tunnel breakdown (see Subsec. 3.2.7).

Since the current I_1 is specified, an increase in output current is attended by a reduction in diode current. With the output ter-

¹ The minus sign on the right of Eq. (9.78b) means that an increase in output current ($\Delta I_2 > 0$) is accompanied by a decrease in output voltage ($\Delta V_2 < 0$).

minals short-circuited ($R_l = 0$), we get $I_d = 0$; in other words, the diode regulator “does not fear” short circuiting at the output. This feature is inherent in all the regulators of the *shunt type*, in which a regulating element is connected in parallel with the load resistance.

The output resistance of a diode regulator, according to Eq. (9.80), is

$$R_{out} = r_d \parallel R_i \approx r_d \quad (9.81)$$

The value of r_d commonly ranges from 10 to 20 Ω at a rated current of 5 to 10 mA; it somewhat rises with decreasing current.

In the case of an ideal current source ($R_i = \infty$), the stabilization factor extends to infinity because the changes in input voltage

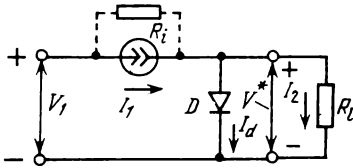


Fig. 9.30. Diode regulator with a forward-biased diode

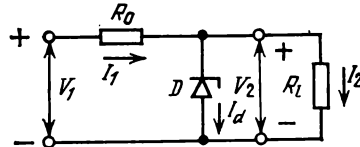


Fig. 9.31. Diode regulator with a ballast resistor

ΔV_1 , in no way cause changes in the output. With R_i being a finite value, the increments ΔV_2 and ΔV_1 are related by the transfer ratio of a resistance voltage divider (see Fig. 9.29b):

$$\Delta V_2 = \Delta V_1 \frac{r_d}{R_i + r_d} \approx \Delta V_1 \frac{r_d}{R_i}$$

Substituting the ratio $\Delta V_2/\Delta V_1$ into Eq. (9.79) yields

$$K_{st} = \frac{V_2}{V_1} \frac{R_i}{r_d} \quad (9.82)$$

For example, if $V_2/V_1 = 0.8$, $R_i = 50 \text{ k}\Omega$, and $r_d = 10 \text{ }\Omega$, then $K_{st} = 4000$.

This value of K_{st} is quite acceptable in most cases. What presents the main problem is a comparatively large output resistance of the regulator.

A reference diode (a diode relying on the breakdown effect) can be replaced by a **forward-biased** diode (Fig. 9.30). In this case the output voltage V_2 is equal to V^* , and the output resistance

$$R_{out} = r_d = (\varphi_T/I_d) + r_b \quad (9.83)$$

where the augend is an incremental junction resistance given by Eq. (3.25), and the addend is the base layer resistance. As the load current rises, I_d decreases and the output resistance substantially

increases. A minimum value of R_{out} corresponds to the open-circuit condition ($I_2 = 0$), but it is always limited by the value of r_b (usually not less than 2 to 5 Ω).

Substituting r_d from Eq. (9.83) into (9.82) and $V_2 = V^*$ gives

$$K_{st} = \frac{V^*}{V_1} \frac{R_i}{\phi_T/I_d + r_b} \quad (9.84)$$

For example, if $R_i = 50 \text{ k}\Omega$, $I_d = 1 \text{ mA}$, $r_b = 5 \text{ }\Omega$, and $V_1 = 5 \text{ V}$, then $K_{st} \approx 220$ and $R_{out} = 30 \text{ }\Omega$.

As compared with the parameters of the preceding circuit, the output resistance here is three times as large and K_{st} is by a factor of 20 smaller.

In both circuits discussed, the current source can be replaced by a *ballast* resistor (Fig. 9.31). In this circuit version, the output resistance is practically the same as in the above circuits, but the stabilization factor has a distinctive feature due to the interrelation between the input and output voltages:

$$V_1 = V_2 + I_1 R_0$$

Substituting V_1 into Eq. (9.82) and replacing R_i by R_0 yields

$$K_{st} = \frac{V_2}{V_2 + I_1 R_0} \frac{R_0}{r_d} \quad (9.85a)$$

As seen in this version K_{st} approaches a finite value as R_0 rises to infinity:

$$K_{st \text{ max}} = V_2/I_1 r_d \quad (9.85b)$$

Assume V_2 and r_d are the same in value as in the preceding example (0.7 V and 30 Ω) and $I_1 = 2 \text{ mA}$. Then $K_{st \text{ max}}$ would be near 12, which is much below the value for the circuits using a current source.

The temperature drift in output voltage of diode regulators is a function of the temperature sensitivity of diodes. Combining the diodes operating in the avalanche region (for which $\text{TCV} > 0$) with diodes operating in the forward bias region (for which $\text{TCV} < 0$) permits decreasing the temperature coefficient to 0.01 % $^{\circ}\text{C}^{-1}$ and below.

9.10.3. Transistor regulators. A schematic diagram of the simplest transistor regulator appears in Fig. 9.32a, and its small-signal model in Fig. 9.32b. The regulator has the configuration of an emitter follower: the load is connected to the emitter circuit and the base is fed with a dc *reference voltage* V_{ref} rather than with an ac signal. The reference voltage source is commonly a diode regulator.

The inspection of the circuit indicates that $V_2 = V_{ref} - V^*$, that is, the output voltage is defined by the reference voltage.

Disregarding the base current, the input and output currents become practically equal: $I_c \approx I_2$. Consequently, an increase of

load current causes the same increase of collector current and, hence, of power dissipated in the transistor. Obviously, short circuiting at the output may lead to overloading of the transistor. This conclusion applies to all transistor regulators of the *series type*, in which the regulating element is connected in series with the load. Series-type power regulators are furnished with a special overload protection system.

The output resistance of this regulator is the same as in the emitter follower. Considering that the role of R_g is played here by a small

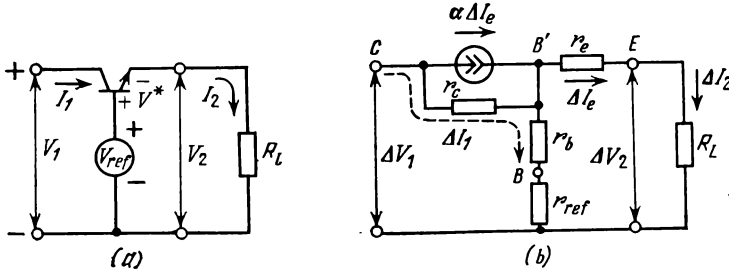


Fig. 9.32. Schematic diagram (a) and circuit model (b) of the transistor regulator using an emitter follower

resistance of the reference element, r_{ref} , it is permissible to use expression (9.64):

$$R_{out} = r_e \quad (9.86)$$

Hence, the output resistance grows with decreasing current, and, as the regulator approaches the open-circuit condition, this resistance may reach unacceptable values. To smooth out the dependence of output resistance on the load current, a fixed resistor can be connected across the load (ahead of the output terminals). This shunt will secure a certain residual emitter current even in the open-circuit condition (at $I_2 = 0$).

If we neglect the collector junction resistance, that is, set $r_c = \infty$, then the increments ΔV_1 will fail to reach the base and emitter circuits of the transistor because the collector circuit includes an ideal current source (Fig. 9.32b). In this case, $K_{st} = \infty$.

Considering the resistance r_c , the increment ΔV_2 versus ΔV_1 may be written through the following relations (see Fig. 9.32b)¹:

$$\Delta V_2 = \Delta V_{b'} \frac{R_L}{R_L + r_e} \approx \Delta V_{b'}$$

$$\Delta V_{b'} = \Delta V_1 \frac{r_b + r_{ref}}{r_c + r_b + r_{ref}} \approx \Delta V_1 \frac{r_b + r_{ref}}{r_c}$$

¹ The current source $\alpha \Delta I_e$ is considered inactive (absent) because the increment ΔI_e is known to be very small (since $V_2 \approx \text{constant}$).

where the right sides contain the transfer ratios of the respective resistance voltage dividers.

Substituting the ratio $\Delta V_2/\Delta V_1$ in Eq. (9.79) gives

$$K_{st} = \frac{V_2}{V_1} \frac{r_c}{r_b + r_{ref}} \quad (9.87)$$

Setting $V_2/V_1 = 0.8$, $r_c = 250 \text{ k}\Omega$, $r_b = 150 \text{ }\Omega$, $r_{ref} = 20 \text{ }\Omega$, we get $K_{st} \approx 1200$.

Thus a *simple single-transistor regulator provides a sufficiently high stabilization factor but does not enable a principal solution of the problem of decreasing the output resistance*. The way of handling this problem today with the use of an op-amp circuit will be discussed in Subsec. 10.10.5.

9.11. Current Regulators

In the above discussion we have repeatedly used current sources as conventional elements of electric circuits. Current regulators described in this section represent the real embodiment of these sources in circuit form.

The task of a current regulator is to maintain a fixed output current I_2 with changes in the load R_L and input (supply) voltage V_1 ,

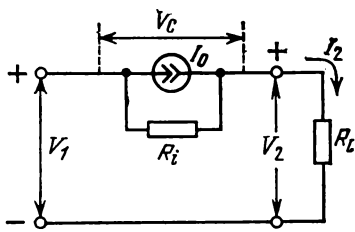


Fig. 9.33. Skeleton diagram of a current regulator

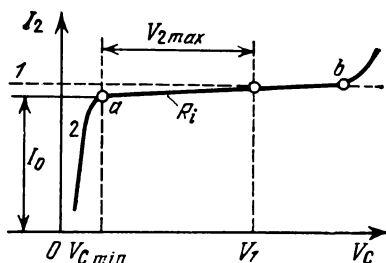


Fig. 9.34. I - V characteristic of a current regulator

(Fig. 9.33). The voltage across a regulating element will further be designated by V_C .

9.11.1. Regulator parameters. An ideal current regulator has a current-voltage characteristic as shown in Fig. 9.34 by a dash line 1. The real regulator characteristics (curve 2) differ from the ideal by a limited working region and a finite incremental resistance R_i in the working region. The resistance R_i generally depends on voltage V_C . But in practice it is unjustifiable to take into consideration such a nonlinearity. Therefore, R_i is always regarded as a certain averaged value.

The points a and b at which R_i drops significantly define the limits of the working region. The point a for bipolar transistors is the limit beyond which they begin to saturate. For MOSTs, this point corresponds to the transition into the steep portion of the characteristic. The point b for both transistor types represents a limit beyond which breakdown sets in.

The voltage $V_{C \min}$ at point a is a minimum voltage across the regulating element that is still sufficient for the regulator to perform its function. In bipolar npn transistors, the minimum voltage may be considered to be equal to 0.2 V for the CE configuration and to 0.5 V for the CB configuration (see Figs. 4.14a and 4.15a). In MOSTs, the minimum voltage is close to a saturation voltage $V_{d \text{ sat}}$ (see Fig. 5.7a). Depending on the current and specific transconductance of a transistor, $V_{d \text{ sat}}$ lies in the range from fractions of a volt to a few volts.

In most electronic circuits, the load of current regulators is nonlinear. For nonlinear elements the resistance is only a symbol identifying the relation between voltage and current. It is, therefore, more convenient to characterize a nonlinear output circuit by the output voltage V_2 at a given current rather than by R_l .

The **nominal** condition of operation of a current regulator is a short-circuit condition at which $V_2 = 0$. Here, as seen from Fig. 9.33, a maximum voltage equal to V_1 drops across the regulating element. As a matter of fact, V_1 must lie within the working region (see Fig. 9.34).

The short-circuit current is considered a nominal current of the regulator. This current is the sum of I_0 and V_1/R_i , the latter being conditioned by the finite resistance of a regulating element (see Fig. 9.33). In all practical regulators, the second component is negligible, so that the rated output current I_2 is assumed to be equal to the current I_0 of an ideal regulating element.

In the condition different from the short-circuit one, the regulator has a finite voltage V_2 at its output. The higher the output voltage, the lower the voltage V_C across the regulating element. The maximum voltage across the load corresponds to a value of $V_{C \min}$.

The quantities I_0 , V_1 , and V_2 are in general uncorrelated, that is, they change independent of one another. In the worst case, therefore, the increment in output current is the arithmetic sum of three increments (see Fig. 9.33):

$$\Delta I_2 = \Delta I_0 + \frac{\Delta V_1}{R_i} + \frac{\Delta V_2}{R_l}$$

Dividing the left side by I_2 and the right side by I_0 which is close to the former in value, we write the relative instability in the form

$$\frac{\Delta I_2}{I_2} = \frac{\Delta I_0}{I_0} + \frac{\Delta V_1}{E_i} + \frac{\Delta V_2}{E_l} \quad (9.88)$$

Here $E_i = I_0 R_i$ is the equivalent voltage determining the degree of stabilization with respect to output and input voltage changes. The higher the value of E_i , the higher the degree of stabilization.

In practice, it is transistors that play the role of regulating elements. For bipolar transistors connected in the CB configuration, I_0 is a collector current and R_i is a collector junction resistance r_c . As is apparent from Eq. (4.42), r_c is inversely proportional to I_c . Their product, therefore, that is, the voltage E_i , is independent of current and can be regarded as a transistor parameter. This parameter is typically 2 or 3 kV. For a CE configuration, in which $r_c^* = r_c (1 - \alpha)$, E_i is correspondingly much lower.

For MOS transistors, the product $I_d r_d$ according to Eq. (5.20) is also a parameter. The typical values of E_i range from 200 to 500 V.

Many circuits use a resistor as a regulating element (see Fig. 8.15). In the analysis of this circuit version, I_0 given in Fig. 9.33 should be taken equal to zero. Then

$$I_2 = (V_1 - V_2)/R_i$$

To have the stabilized current weakly dependent on the output voltage, the condition $V_2 \ll V_1$ must be met; in this case, $I_2 \approx \approx V_1/R_i$.

The relative instability assumes the form analogous to that of Eq. (9.88):

$$\frac{\Delta I_2}{I} = \frac{\Delta V_1}{V_1} + \frac{\Delta V_2}{V_2}$$

The supply voltage V_1 does not generally exceed 10 to 15 V, that is, it is by far smaller than the equivalent voltage E_i in nonlinear regulating elements [see Eq. (9.88)]. Other things being the same, the current instability then will be much higher.

9.11.2. Simple regulators. Fig. 9.35a illustrates a typical circuit form of the current regulator where E_0 is the source of a stabilized voltage. In the given case, the source uses a reference diode fed via a ballast resistor (see Fig. 9.34).

Figure 9.35b shows the regulator circuit model where heavy lines represent an output section. The designations of the basic quantities are the same as in Fig. 9.33; the element L identifies a nonlinear load. Comparing the output circuit of Fig. 9.35b with the circuit of Fig. 9.33, we find

$$I_0 = \alpha I_e = \alpha \frac{E_0 - V^*}{R_0} \quad (9.89a)$$

$$R_i = r_c \quad (9.89b)$$

$$V_1 = E_c + E_e - E_0 \quad (9.89c)$$

Expressions (9.89) permit us to make the following conclusions.

The stabilized current is set by the components E_0 , R_0 . The stability of current is primarily determined by the stability of E_0 and V^* . In particular, if the temperature drifts of these quantities are opposite in sign, then the temperature drift of current will be in excess of each of the above drifts.

The internal resistance R_i rises with decreasing current [see Eq. (4.42)]. The value of R_i given by Eq. (9.89b) corresponds to the

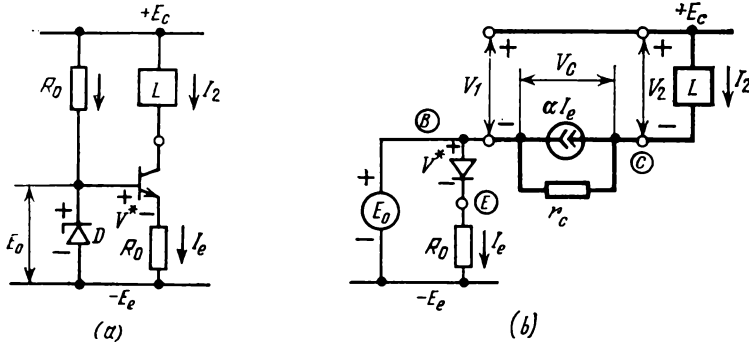


Fig. 9.35. Schematic diagram (a) and circuit model (b) of the simple current regulator based on an npn transistor

absolutely invariable emitter current. In practical circuits, where R_0 is a finite value, the increment ΔI_2 is distributed between the emitter and base circuits. The current increment in the emitter circuit is

$$\Delta I_e = \gamma_e \Delta I_2$$

where γ_e is given by (9.14). The role of R_g in the given case is played by the resistance of reference diode D in Fig. 9.35a. The increment in output current is thus the sum of two components:

$$\Delta I_2 = \Delta V_2 / r_c + \alpha \gamma_e \Delta I_2$$

From this expression it is easy to obtain the output resistance in the general form

$$R_i = \Delta V_2 / \Delta I_2 = r_c (1 - \alpha \gamma_e) \quad (9.90)$$

The maximum output voltage in the given circuit can lie near V_1 given by Eq. (9.89c) since the transistor retains its amplifying properties even at V_{cb} close to zero.

The instability of current I_2 may be estimated by general formula (9.88) using expressions (9.89).

If the load is connected not to a positive but to a negative supply source or to "ground", then the regulator is built around a *pnp*

transistor (Fig. 9.36a). This circuit is similar to the preceding one and described by the same expressions.

In practical use is sometimes the type of regulator as shown in Fig. 9.36b. What distinguishes this regulator type is single-polarity power supply and zero potential of the base (disregarding a small voltage drop $I_b r_b$). As a result, the collector potential proves positive with respect to the base potential, that is, the collector junction turns out to be **forward biased**. The transistor formally operates in the double injection mode. As known, however, a small forward

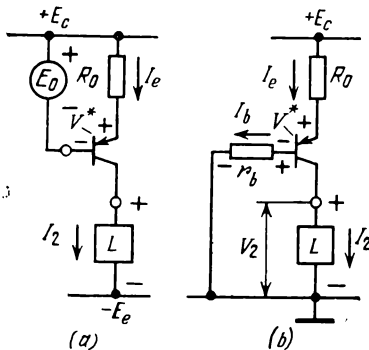


Fig. 9.36. Simple current regulators based on a *pnp* transistor
(a) basic circuit version; (b) circuit version with the base grounded

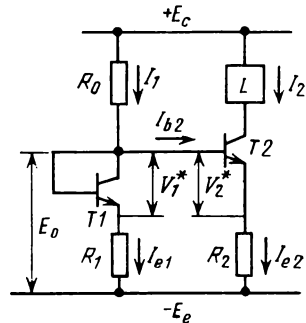


Fig. 9.37. Current reflector

bias (below $V^* - 0.1$ V) on the collector junction does not cause a noticeable reduction in collector current (see curves in Fig. 4.14a in the second quadrant). Consequently, over the range of rather low output voltages (tenths of a volt), the given regulator works normally.

9.11.3. Current reflectors. Analog integrated circuits widely use a current regulator known as current reflector or current "mirror" (Fig. 9.37). It is easy to see a formal resemblance between the current reflector and simple regulator circuit: instead of the reference diode (see Fig. 9.35a), the circuit employs a resistor R_1 and forward-biased *pn* junction, the role of the latter being played by a transistor *T1* connected in the BC-E circuit to perform a diode action (see Fig. 7.24). Such a version of the source E_0 leads to an increased flexibility of the circuit and improvement in the number of parameters.

The inspection of Fig. 9.37 indicates that

$$V_1^* + I_{e1}R_1 = V_2^* + I_{e2}R_2 \quad (9.91)$$

This equality underlies the operation of the current reflector.

Resistance-containing summands in Eq. (9.91) do not generally exceed the value V^* . So, depending on the working currents, resistances R_1 and R_2 range from hundreds of ohms to 10 to 20 kilohms.

Disregard for simplicity the small current I_{b2} ; hence, $I_{e1} = I_1$ and $I_{e2} = I_2$. Besides, suppose that R_1 and R_2 are equal and $T1$ and $T2$ are identical (in the IC, it is easy to make transistors identical since the elements on the chip lie close to each other). Given such conditions, the summands on the right and the left of the equality (9.91) will be the same, and hence $I_2 = I_1$.

Thus, in the discussed circuit version, the output current repeats or reflects the input current I_1 . Hence, the name current reflector.

The input current, as follows from Fig. 9.37, takes the form

$$I_1 = (E_c - E_0)/R_0$$

If $E_c \gg E_0$, then I_1 is governed by the parameters E_c and R_0 . In many cases, the current I_1 flows from certain stages forming part of a complex electronic system, and thus can be regarded as a specified value.

If R_1 and R_2 are made unequal, the emitter currents will be unequal too. Since V_{eb} is weakly dependent on current, assume as before that $V_1^* = V_2^*$. From Eq. (9.91) it then follows that

$$I_2 = I_1 (R_1/R_2) \quad (9.92)$$

As seen, the current I_2 can "reflect" I_1 both in an "enlarged" and in a "reduced" scale. This scale does not commonly exceeds a few units, otherwise a resistor of high rating occupies too large an area.

From Eq. (9.92) we can conclude that *the output current I_2 can be controlled* by varying the input current I_1 in any of the suitable ways. This possibility is one of the features indicative of the flexibility of a current reflector.

If we allow for the current I_{b2} , then the emitter current will not be exactly equal to currents I_1 and I_2 ; namely,

$$I_{e1} = I_1 - I_{b2}, \quad I_{e2} = I_2 + I_{b2}$$

It is then necessary to apply requisite corrections to Eq. (9.92). At high values of B , in which case $I_{b2} \ll I_{e2}$, these corrections are insignificant.

It should be pointed out that Eq. (9.92) does not contain either V^* or B . This means that to a first approximation the operation of a current reflector does not depend on changes in these parameters, primarily on temperature changes. In reality, such a dependence, though rather weak, does exist. The thing is that the gain B will enter into Eq. (9.92) after allowing for I_{b2} . The same also concerns V^* if we take account of the quantity $V_1^* - V_2^*$ resulting from the difference between emitter currents.

To ensure especially small output currents (for example, when a DA operates in the microampere region), the resistance R_1 should be brought down to zero. For this type of current reflector (Fig. 9.38a), formula (9.92) is invalid because the difference between V_1^* and V_2^* cannot be disregarded. Let us use Eq. (4.36b) for V_1^* and V_2^* . Setting

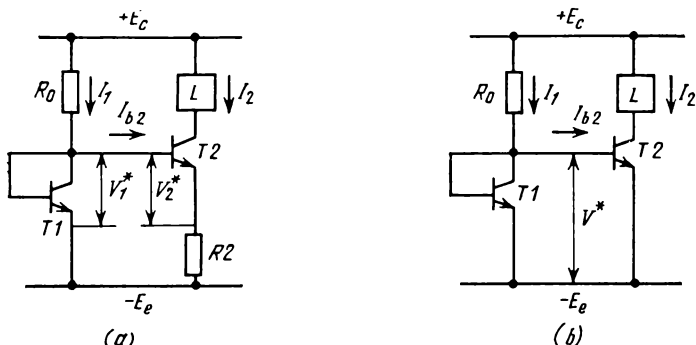


Fig. 9.38. Current reflectors with a single resistor (a) and without a resistor (b)

$R_1 = 0$, we can readily obtain a transcendental relation between the output and input currents from Eq. (9.91):

$$I_2 = (\varphi_T/R_2) \ln (I_1/I_2) \quad (9.93a)$$

A more illustrative expression is the approximation relating the currents in the explicit form:

$$I_2 \approx \sqrt{\left(\frac{\varphi_T}{R_2}\right) I_1} \quad (9.93b)$$

It is clear that in the given circuit the dependence of I_2 on I_1 is substantially weaker than for the preceding circuit; in other words, control of the output current is more difficult.

Expressions (9.93) can be readily applied for the calculation of the required resistance R_2 if the desired values of currents are set.

For example, assume $I_1 = 0.5$ mA and $I_2 = 10$ μ A; formula (9.93a) then gives $R_2 \approx 10$ k Ω . Here the voltage drop $I_2 R_2$ (the difference $V_1^* - V_2^*$) will come to about 100 mV.

Expressions (9.93) show that the currents ratio is independent of the gain B as before, but the current I_2 is directly dependent on temperature through the thermal potential φ_T . To decrease this dependence, it is desirable that the resistor R_2 have the same temperature coefficient as φ_T , namely, 0.33% $^{\circ}\text{C}^{-1}$. These values of TCR are easy to secure for integrated resistors (see Subsec. 7.9.1).

One more type of current reflector (Fig. 9.38b) is distinguished by the absence of resistors and thus noted for its minimum area

on the chip. But if $I_1 \neq I_2$, one of the transistors has to be made larger and so the saving in area offered by this type of circuit over the resistor circuit version is considerably less.

Setting $R_1 = R_2 = 0$ in (9.91) and substituting V_1^* and V_2^* from (4.36b), one may readily see that the ratio I_2/I_1 is proportional to the ratio between thermal currents, I_{e02}/I_{e01} . Other conditions being equal, thermal currents are proportional to junction areas as follows from (3.18). In integrated circuits, "other equal conditions" (that is, identical electrophysical parameters) are achieved by means of a close location of transistors with respect to each other. Summing up all that has been said, for the circuit of Fig. 9.38b we may write

$$I_2 = I_1 (S_2/S_1) \quad (9.94)$$

where S_1 and S_2 are the areas of emitter junctions.

As with the basic circuit version of Fig. 9.37, this circuit is free from the effect of changes in B and V^* to a first approximation. But these changes make itself felt to a definite degree if the current I_{b2} is allowed for (this fact was mentioned above). The larger the area S_2 , and hence the current I_2 , the greater the base current and the larger the error introduced when using Eq. (9.94). In practice, the ratio between the currents and areas rarely exceeds a few units, so the error runs from 2 to 5%.

Note in conclusion that the output resistance R_i in a current reflector may noticeably differ from the resistance r_c taken for a simple regulator [see Eq. (9.89b)]. The thing is that in a simple regulator, the resistance of an emitter circuit is always much higher than that of a base circuit because of a rather large resistance R_e . For this reason the coefficient γ_e in (9.90) does not usually exceed 0.1 or 0.2. In current reflectors, the resistances of emitter and base circuits of a transistor $T2$ may be in various proportions. In particular, the inequality $R_b > R_e$ may prevail. In this case, the output resistance of $T2$ and hence the internal resistance of the current reflector should be calculated by Eq. (9.90).

For example, in the circuit of Fig. 9.38b the resistance R_b should be regarded as r_{e1} and R_e as r_{e2} . According to Eq. (4.41), both of them are inversely proportional to respective emitter currents, practically to currents I_1 and I_2 . Therefore, if we set $\alpha = 1$ and $r_{b2} = 0$ for simplicity, then R_i given by Eq. (9.90) will take the following form after simple transformations:

$$R_i \approx \frac{r_c}{1 + I_2/I_1}$$

As seen, with a rise in the ratio I_2/I_1 the resistance R_i substantially decreases. This conclusion is general enough to be valid for other discussed circuits of current reflectors.

10

10.1. General

In the preceding chapters we have familiarized ourselves in detail with the basic physical, design, manufacturing, and circuit engineering concepts of microelectronics. In the concluding chapter we shall consider examples of simple ICs and some problems involved in their development.

The first integrated circuits that appeared in 1960 were simple digital ICs based on bipolar transistors. From the middle of the 1960s, the development of bipolar analog ICs and MOS digital ICs began. The first half of the 1970s was marked by the appearance of some new circuit designs specific to microelectronics (charge-coupled devices, I²L circuits, and others). This period also saw a sharp rise in the scale of integration and establishment of a kind of “dynamic equilibrium” between the basic classes of ICs—bipolar and MOS transistor ICs and also monolithic and hybrid ICs. Before then, the above classes were frequently regarded as alternative. From the second half of the 1970s onward, the scale of integration grew still more rapidly, and sophisticated devices appeared in integrated form capable of performing complex and numerous operations.

A detailed description of LSI circuits—their structure, design principles, available types, applications, etc.—is not only impossible within the scope of the present book, but also hardly advisable. For this reason, the text below only gives an idea of the “building blocks” which make up modern digital ICs and largely determine their performance. As for analog ICs, we shall consider an example of the op amp which is best suited for revealing their features.

10.2. Bipolar Logic Elements

Logic elements, or *logic gates*, are electronic circuits performing simple logical operations. Before getting into the details of the circuit versions of logic elements let us dwell on the functions they have to perform.

10.2.1. Logic functions. Logic functions and the logical operations involved relate to *Boolean algebra*. What underlies Boolean algebra is the system of logical variables which will be designated as A , B , C , etc. A logical variable characterizes two incompatible

notions: yes and no, black and not-black, turn-on and turn-off, etc. If one of the values of the logical variable is denoted as A , then the second is denoted as \bar{A} (not A).

In dealing with logical variables, it is convenient to use a binary code, assuming $A = 1$, $\bar{A} = 0$, or, inversely, $A = 0$, $\bar{A} = 1$. The same circuit can thus perform both logical and arithmetic operations (in the binary number system).

If we denote "not A " by a certain letter, B for example, then the relation between the values of B and A will take the form

$$B = \bar{A} \quad (10.1)$$

This is a simple logic function known as *negation*, *inversion*, or *NOT function*. The circuit capable of performing this function is known

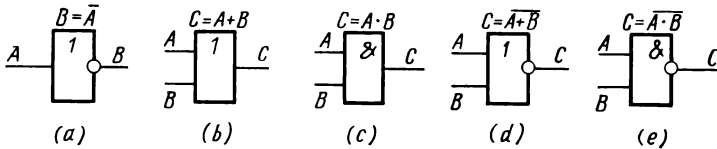


Fig. 10.1. Logic symbols for gates
(a) NOT; (b) OR; (c) AND; (d) NOR; (e) NAND

as an inverter, or NOT gate, whose symbol is shown in Fig. 10.1a. The NOT function is characterized by a circle on the output side of the rectangle.

The NOT function is the function of **one** argument (one variable). Let us give example of logic functions for two variables.

Logical addition, disjunction, or OR function:

$$C = A + B \quad (10.2)$$

This function is defined in the following manner: $C = 1$ if $A = 1$ or $B = 1$, or both $A = 1$ and $B = 1$. The symbol for an OR gate is shown in Fig. 10.1b.

Logical multiplication, conjunction, or AND function:

$$C = AB \quad (10.3)$$

This function is defined as follows: $C = 1$ only if $A = 1$ and $B = 1$ simultaneously. The symbol for an AND gate is given in Fig. 10.1c.

The combination of a NOT function and an OR function results in a *NOT OR* or *NOR operation* (Fig. 10.1d):

$$C = \overline{A + B} \quad (10.4)$$

By analogy, the combination of a NOT function and an AND function results in a *NOT AND* or *NAND* operation (Fig. 10.1e):

$$C = \overline{AB} \quad (10.5)$$

The NOR and NAND functions are most widespread because *they permit of implementing any other logic function*. As a matter of fact, the number of variables and hence the number of inputs in corresponding circuits can be equal to three, four, and more.

In circuits intended to implement a logic function, that is, in logic elements, logic 0 and logic 1 are generally represented by different values of voltage: a voltage level V^0 for 0 (logic level 0) and a voltage level V^1 for 1 (logic level 1).

If logic level 1 is more positive than level 0 the circuit is said to use the positive logic convention, and if level 1 is more negative than level 0 the circuit is said to use the negative logic convention. There is no principal difference between positive and negative logic. Moreover, as will be shown later, the same circuit can perform both the positive-logic and the negative-logic operation. Elsewhere below we shall assume that the circuits perform the positive-logic operation, which is in accord with its popularity in practice.

The difference between the voltage levels for 1 and 0 is termed the *logic swing*:

$$V_l = V^1 - V^0 \quad (10.6)$$

The logic swing must naturally be large enough so that logic 1 and logic 0 should differ sharply and no parasitic signal could "convert" one level into another.

Integrated logic elements form the basis or, what is termed, *building blocks* for more complex ICs and systems as a whole. The parameters of IC logic elements affect directly the parameters of units and subsystems. In other words, the choice of the IC logic element type largely predetermines the performance of equipment.

IC logic elements in circuit form are usually called *transistor logic circuits* (though the term is not very fitting) and denoted by TL, with the addition of appropriate symbols that characterize a particular circuit version. But this system of notation is not strictly adhered to.

10.2.2. Direct-coupled transistor logic (DCTL) and its versions. IC logic elements of this logic family basically use conventional parallel-connected transistor switches with a common collector load (Fig. 10.2).

The simplest and historically first version of these integrated DCTL elements is shown in Fig. 10.2a by solid lines. Dash lines represent the transistors entering into other similar IC logic elements. The preceding logic element represented by a transistor T_3

controls a switch $T1$, and transistors $T4$ and $T5$ that enter into the next logic elements are the load for the given element.

It is easy to see that the circuit performs the NOR function when operated in positive logic. Indeed, if a low voltage level $V^0 \approx 0$ is

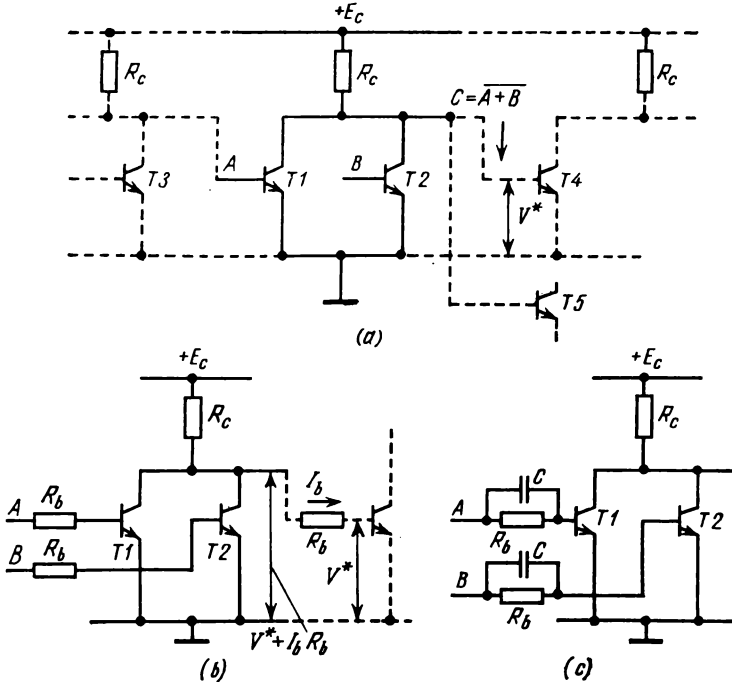


Fig. 10.2. Logic elements of the DCTL family

(a) DCTL circuit; (b) RTL circuit; (c) RCTL circuit

applied to inputs A and B , both transistors stay off, the current passes through a resistor R_c into the bases of $T4$ and $T5$, so that the output voltage of the logic element (see Fig. 10.2a) is

$$V^1 = V^* \quad (10.7a)$$

But if a high voltage level $V^1 = V^*$ is provided at one of the input terminals of the logic element, the appropriate transistor becomes biased on and saturated at a sufficiently large base current. The output voltage level is then low and equal to a residual voltage

$$V^0 = V_{res} \quad (10.7b)$$

The same level results if both transistors switch on. So formula (10.4) holds here.

In the negative-logic operation ($V^0 = V^*$, $V^1 = V_{res}$), the circuit acts as a NAND gate: a high output level (V^0) is obtained only if both transistors are held in cutoff by applying low levels (V^1) to both inputs. *A change in the nature of the function implemented on converting from positive logic to negative logic, and vice versa, is a property common to all the IC logic elements, and so further we shall not illustrate this feature for each circuit.*

Taking into account the above values of logic levels, the logic swing for a DCTL circuit may be written as

$$V_l = V^* - V_{res} \quad (10.8)$$

For the normal current region and microampere region, V_l is respectively equal to 0.6 V and 0.5 V.

Since the logic swing is low over the supply voltage (which is commonly 3 to 5 V), the changes in current through the collector resistor R_c prove insignificant, that is, the *current remains almost constant*. It thus can be said that in a DCTL circuit, a change in the output level from V^0 to V^1 brings about only **redistribution of current** from the collector circuit of the given logic element into the base circuits of loads.

The relations required for circuit design, in particular, for the calculation of a circuit load capability, can be taken from Sec. 8.3.

A serious disadvantage of DCTL is a nonuniform distribution of current between the bases of load transistors. The cause is the difference between input I - V curves (see Fig. 8.5b). This difference by no means arises from the spread in parameters due to manufacturing factors (which is very small in ICs), but primarily results from the inevitable difference in the collector currents of saturated load transistors. The dependence of the input I - V curve on saturation current follows from Eq. (8.6). For example, if the logic element of Fig. 10.2a has only one of the transistors turned on, then its saturation current is equal to E_c/R_c ; if both transistors are on, the current in each is $1/2 E_c/R_c$.

The nonuniform distribution of a load current makes the DCTL circuit operation unreliable. This circumstance explains why this form of logic has evolved into other, more dependable forms.

One of these forms—*resistor-transistor logic* (RTL)—is shown in Fig. 10.2b. It differs from DCTL in that the base circuits of transistors contain resistors of a few hundred ohms. These resistors permit equalizing the input I - V characteristics (see Fig. 8.5b), and hence balancing out base currents. The logic 1 level and logic swing here grow to $V^1 = V^* + I_b R_b$ (typically 1.5 to 2 V), but the base current turns out to be smaller than for the DCTL circuit. A decrease in the base current lowers the load capability and reduces the speed of the RTL circuit as compared to the DCTL circuit because of the

increase in the rise time of a signal in the transistor switch [see Eq. (8.17)].

A second form of DCTL—*resistor-capacitor-transistor logic* (RCTL)—is shown in Fig. 10.2c. This logic circuit differs from the preceding one by the presence of small-value capacitors shunting the resistors. At the moment of switching of the preceding logic element, these

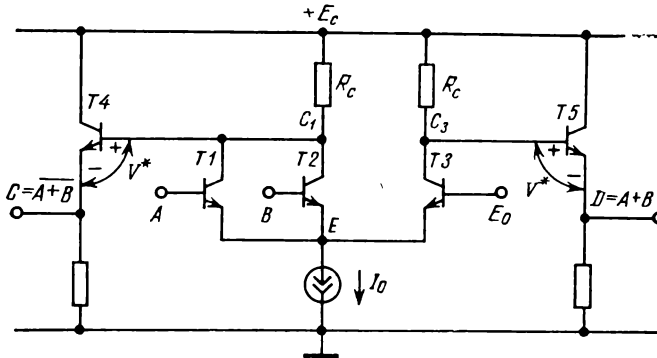


Fig. 10.3. ECL gate circuit

capacitors “short out” the resistors for some time and ensure the increased values of base currents. The positive pulse rise time thus noticeably decreases.

RTL and RCTL circuits were used in the first stage of development of microelectronics. However, in semiconductor ICs of high packing density, they proved to have little promise on account of a considerable number of resistors and capacitors occupying a large area. One more, main, version of DCTL that plays an important role in modern microelectronics is considered separately in Sec. 10.3.

10.2.3. Emitter-coupled transistor logic. A more widespread, though less accurate, term for this form of logic is emitter-coupled logic (ECL).

The ECL circuit (Fig. 10.3) uses a current switch described in Sec. 8.6, for which reason it is sometimes called a current-mode logic (CML) circuit. But in distinction to a simple switch, the circuit here has a few parallel-connected transistors ($T1$ and $T2$ in Fig. 10.3) placed in one of the circuit branches. These transistors are equipotent in the sense that the turn-on condition of any one (or all together) leads to switching of the current I_0 from the right branch to the left. Therefore, this logic circuit, like the DCTL circuit, performs the NOR function.

The emitter followers $T4$ and $T5$ shift the levels of collector potentials by the value V^* . As known, current switches cannot operate

jointly without this level shifting (see Fig. 8.14 where the value of shift is denoted by e). Substituting $e = V^*$ into (8.40) and setting $\delta \approx 0.1$ V, we find that the ECL network is fit to work with the emitter followers present.

Assume the voltages of level V^0 are applied to both logic inputs, V^0 being low enough to make the emitter junctions of $T1$ and $T2$ reverse biased. According to the NOR function, the voltage at the output C will then equal level V^1 . This level is easy to determine considering that $T1$ and $T2$ are off, that is, $V_{c1} = E_c$. Subtracting the voltage V^* across the emitter junction of $T4$ gives

$$V^1 = E_c - V^* \quad (10.9a)$$

Next, apply a forward bias voltage V^1 to one of the inputs, input A for example. The voltage at the output C will not assume a value of V^0 . Suppose, as we did in Sec. 8.6, that the on transistor stays at the **edge** of saturation, that is, $V_{c1} = V_{b1} = V^1$. Subtracting V^* and substituting Eq. (10.9a) yields

$$V^0 = E_c - 2V^* \quad (10.9b)$$

Using Eqs. (10.9), we find the value of logic swing:

$$V_l = V^* \approx 0.7 \text{ V} \quad (10.10)$$

As seen, this value does not practically differ from that typical of the DCTL circuit [see Eq. (10.8)].

The bias voltage E_0 is taken equal to the half-sum of levels V^0 and V^1 :

$$E_0 = E_c - 3/2 V^* \quad (10.11)$$

The levels V^0 and V^1 thus lie symmetric about E_0 , $\pm 1/2 V^*$ distant from it.

The operating current I_0 is estimated from the equality $V_{c1} = V^1$ that ensures the boundary conditions for an on transistor. Substituting $V_{c1} = E_c - \alpha I_0 R_c$ and V^1 given by Eq. (10.9a) into this equality yields

$$\alpha I_0 R_c = V^* \quad (10.12)$$

The resistance R_c is found from the condition (8.44) that provides for a minimum switching time. The values of R_c usually lie from 0.5 to 2 k Ω , and those of I_0 from 0.35 to 1.5 mA. In calculating the power consumed, it is certainly necessary to allow for the currents in emitter followers.

Using Eqs. (10.9b) and (10.11), it is easy to show that in the initial state, with the level V^0 being set at **both** inputs, the emitter junctions of $T1$ and $T2$ are at a **forward** bias of $+1/2 V^* = 0.35$ V. This bias voltage is lower than the turn-off voltage, which is equal to about 0.6 V (see p. 89). The transistors thus **practically** stay off. However, the circuit noise immunity, defined as the difference between the bias

voltage and the turn-off voltage, comes to merely 0.25 V. This value is about half as large as that for a DCTL circuit, where the transistors are driven into cutoff at $V_{res} \approx +0.1$ V, which is 0.5 V lower than the turn-off voltage.

To improve the noise immunity of a circuit, it may be found expedient to have the on transistors operated not at the boundary of saturation but in the quasisaturation region, that is, at a low (0.2 to 0.3 V) forward bias on the collector junction. It is also possible to replace the followers by more complex level shifting circuits (see Fig. 9.20). The use of the latter is equivalent to an increase in V^* . But such circuits contain additional pn junctions or resistors and, hence, occupy a large area.

So far it has been assumed that the transistor $T3$ performs an auxiliary function, that is, "stores" the current I_0 when the logic transistors are off. However, it can be readily seen that $T3$ together with the follower $T5$ also implements a logic function; that is, in the initial state, when $V_A = V_B = V^0$, $T3$ remains on, and hence the output voltage V^0 appears at D . If at least one of the logic inputs has a voltage level V^1 , $T3$ does not conduct and thus $V_D = V^1$. Obviously, the ECL circuit provides simultaneous NOR or OR outputs at terminals C and D respectively. That is why the ECL circuit is said to be adapted to realize the *OR/NOR* function.

Let us note in conclusion that in practical ECL circuits it is usual to ground not a negative but a positive terminal of the power supply, and so all the operating voltages in the circuit of Fig. 10.3 are negative. This does not certainly change the principle of the circuit and basic relationships, but grounding of the positive power line greatly decreases the effect of parasitics in the line on the values of levels V^0 and V^1 .

10.2.4. Diode-transistor logic (DTL). In the DTL circuit shown in Fig. 10.4, diodes $D1$ and $D2$ perform a logic function, and a transistor T acts as an inverter. In contrast to the above-described circuits, therefore, in this circuit *the number of transistors is not related to the number of logic inputs*.

Diodes $D3$ and $D4$ do not implement logic functions. Their task is to provide a constant voltage (level shift) between points a and b , for which reason they are called *level shift diodes*. So that the operation of these diodes will be independent of the transistor state (pre-

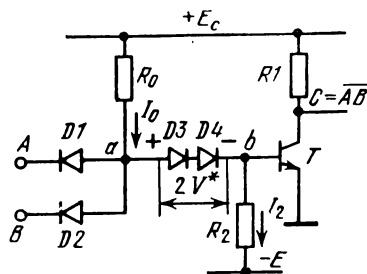


Fig. 10.4. DTL gate circuit

sence or absence of the base current), the DTL logic gate has a level shifting network (R_2 and $-E$). A certain "guard" current I_2 flows through this network, so D_3 and D_4 always operate in the forward direction and provide the shift level $2V^*$.

Let the voltages at logic inputs be equal to zero in the initial state: $V_A = V_B = V^0 = 0$. The diodes D_1 and D_2 are then in the on condition, with I_0 passing through each. The voltage at point a is equal to the forward voltage on the diode:

$$V_a = V^*$$

The current I_0 assumes the form

$$I_0 = (E_c - V^*)/R_0$$

The base potential on T will be $2V^*$ lower than the potential at the point a , that is,

$$V_b = -V^*$$

Hence, the emitter junction of T is reverse biased and the transistor is fully driven into cutoff. This feature is one of the important advantages of the DTL circuit over the circuits discussed earlier, in particular, over the ECL circuit. Note that if the circuit uses only one diode instead of two, the base potential on the off transistor will be near zero, that is, about the same as in the DCTL circuit and its versions.

The level shifting current I_2 is given by

$$I_2 = [V_b - (-E)]/R_2 = (E - V^*)/R_2$$

The resistance R_2 is chosen to be sufficiently large, and so the "guard" current I_2 does not exceed 0.1 or 0.2 mA.

When T is off, the output voltage is a maximum, equal to the supply voltage. This is obviously the logic 1 level:

$$V^1 = E_c \quad (10.13a)$$

Let the input A now be at logic 1 (V^1). Since the voltage at the second input has remained equal to zero, D_2 is on as before, and hence the voltage at point a is still equal to V^* . Consequently, D_1 is driven into cutoff at a large reverse voltage and the current through it drops close to zero. These are the only events that occur in the circuit, and the output is at the level E_c as before. The same will take place if the voltage V^1 is applied to the input B only.

The voltage V^1 applied to both logic inputs reverse biases diodes D_1 and D_2 , and the current I_0 flows the other way—through the level shift diodes to the transistor base. The transistor then switches on, the base potential becomes equal to V^* and the collector potential (in the saturated state only) to V_{res} . This last value is logic

level 0:

$$V^0 = V_{res} \approx 0.1 \text{ V} \quad (10.13b)$$

Thus the output voltage level changes from V^1 to V^0 only when all the inputs are at the level V^1 . So, the DTL circuit in the **positive-logic** operation acts as a NAND gate. This is one more feature of the given circuit that distinguishes it from the above-described circuits performing the NOR function. But the DTL circuit operated in negative logic also realizes the NOR function (see p. 378).

It should be noted that the DTL circuit shows a large logic swing. In a practical circuit, where each integrated logic element is connected to the preceding and the succeeding element, the swing will be the same as in the isolated circuit. Substituting Eqs. (10.13) in Eq. (10.6) gives

$$V_l = E_c - V_{res} \approx E_c \quad (10.14)$$

10.2.5. Transistor-transistor logic (TTL, or T²L). For all its merits

noted above, the DTL circuit suffers from a serious disadvantage that it has a large number of diodes. Because each diode (in essence, this is a transistor connected in the diode configuration) needs an isolated well, the area of an IC logic gate proves rather large.

From Fig. 10.4 it is seen that the combination of logic diodes and level shift diodes corresponds to a transistor structure: two opposite-connected *pn* junctions.

This suggests an idea of replacing the combination of diodes by a multiple-emitter transistor fabricated on one isolating well. The TTL circuit of Fig. 10.5 is the embodiment of this idea.

The TTL circuit naturally performed the same function as the DTL circuit, namely the NAND function. There are two basic features that distinguish one circuit from the other.

First, in the TTL circuit the multiemitter-transistor has only one collector junction that replaces two level shift diodes specific to the DTL circuit. Hence, with logic level 0 kept at the inputs, the base voltage of *T1* will not be negative, as is the case for the circuit of Fig. 8.4, but close to zero (see p. 382)¹. This does not affect the output level V^1 because *T1* remains **practically** off, but impairs somewhat the noise immunity of the circuit. However, the absence of bias source

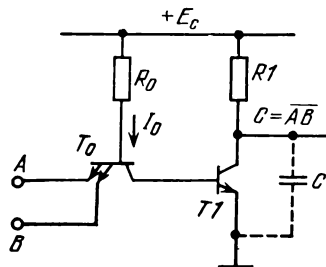


Fig. 10.5. TTL gate circuit

¹ Indeed, the emitter junctions of the on multiemitter transistor are at a forward voltage V^* . The collector voltage of the multiemitter transistor, according to Eq. (4.32b), is also near V^* if the emitter current is present and the collector current is zero. So the difference between V_c and V_e is near zero.

— E and savings in the area for the level shift diodes and resistor $R2$ pay for the impaired noise immunity.

Second, in contrast to isolated diodes, the emitters of a multi-emitter transistor may *interact with each other*, producing what is called the lateral transistor effect (see Subsec. 7.4.1). As a result, the emitter being reverse-biased by V^1 may let pass a parasitic reverse current caused by the injection of electrons from the adjacent, forward-biased emitter. *As it goes through resistor $R1$ in the preceding logic element, this parasitic current will decrease the level V^1 .*

To eliminate the transistor effect, one has to take special precautions (see Subsec. 7.4.1) which generally lead to an increase in the

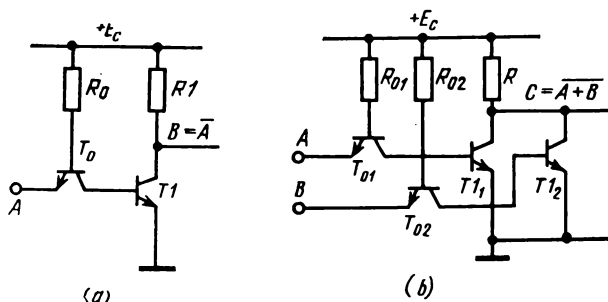


Fig. 10.6. TTL circuit versions
(a) inverter circuit; (b) NOR gate circuit

chip area. All the same, the use of this circuit enables substantial savings in area on the chip as against the DTL circuit. The TTL circuit thus exhibits the space-saving feature and also retains the general advantages of the DTL circuit. That is why the former has practically ousted the latter and presently enjoys the most extensive use. Expressions (10.13) and (10.14) presented in the above subsection are fully valid for TTL circuits.

Figure 10.6a shows a simple inverter using a single-emitter transistor, and Fig. 10.6b shows a NOR gate circuit using two parallel-connected inverters. These examples illustrate the flexibility and versatility of the basic TTL circuit.

One of the tangible limitations of the TTL circuit shown in Fig. 10.5, is its small load capability: if the circuit operates into several loads, the overall load capacitance (shown by a dash line) and the time constant CR_1 with which this capacitance charges grow accordingly. To speed up the charging of the capacitance and raise the load capability, the useful approach is to replace the simple inverter based on a single transistor by a composite inverter composed of three transistors and a diode (Fig. 10.7). This form of logic received the name TTL-3.

In the composite inverter, transistors T_1 and T_3 are connected in a Darlington circuit (see Fig. 9.1) and thus can be regarded as a unit; they switch on and switch off together. The resistor R_2 is not a principal component: a fraction of I_{e1} branches out into R_2 , thereby reducing I_{b3} and thus the degree of saturation of T_3 . The combination of T_2 and T_3 may be regarded as a variant of the push-pull output stage (see Fig. 9.25) where the transistors work in sequence: T_2 charges the capacitor C (current I_1) and T_3 discharges it (current I_2). The diode D drives T_2 into cutoff with T_3 on and in saturation, and the resistor R_3 limits the current through T_2 and T_3 when T_2 turns on while T_3 is not yet fully off.

Note that the use of the composite transistor T_1 - T_3 in the composite inverter improves the noise immunity to the level specific to DTL. The reason is that the network comprising the multiemitter-transistor collector junction connected in series with the emitter junction of T_1 is equivalent to two level shift diodes in the DTL circuit (see Fig. 10.4). Last years have seen extensive use of Schottky-barrier transistors in TTL (see Subsec. 7.4.3). As distinct from conventional types, these transistors enable a higher switching speed (owing to the absence of saturation). However, a concurrent increase in the residual voltage (level V^0) by 0.2 or 0.3 V causes a respective decrease in the logic swing.

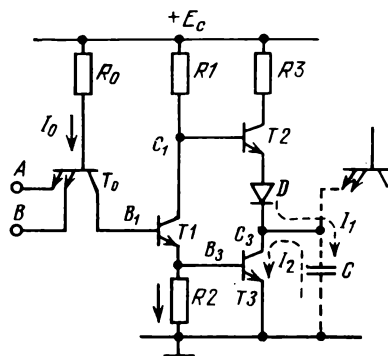


Fig. 10.7. TTL gate circuit using a composite inverter

10.3. Integrated Injection Logic

This new family of IC logic circuits uses an injection-type power source. The principle of injection-type supply is generally applicable not only to digital but also to analog ICs. At present, however, it is practically applied only in digital circuits. Therefore, we consider the principle of injection-type supply using an IC logic element as an example. The adopted term *integrated injection logic* (I^2L) for this form of IC logic does not appear appropriate. Injection-source transistor logic would be more correct and better name for it.

The I^2L circuits are the latest development of IC logic devices. There are no analogs of these circuits in discrete transistor engineering; I^2L units are adaptable for manufacture only in integrated form. In essence, they are another, most improved, modification of the DCTL circuit. It is worth elucidating this continuity because I^2L

circuits appear to be unusual externally and do not remind directly of DCTL circuits.

In I^2L circuits, resistors R_c are changed for dc generators I^* (see Fig. 10.8a and 10.2a). This replacement does not affect the principle of an IC logic element, since in the DCTL circuit the current through R_c practically remains invariable too (see p. 378). The advantages derived from the substitution of a current generator for a resistor

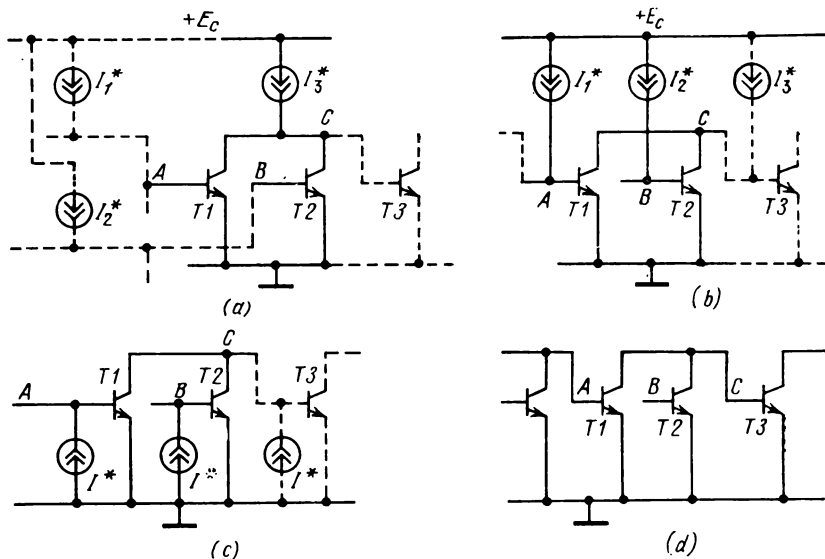


Fig. 10.8. Stages of I^2L gate development

(a) replacement of resistor by current generator in DCTL circuit; (b) shifting of current generators to base circuits of next logic elements; (c) connection of current generator between base and common "grounded" line; (d) I^2L circuit (current generators not shown)

will be explained later. And now let us formally shift all current generators to the "right"—to the bases of load transistors (Fig. 10.8b). This will certainly exert no effect on the circuit operation. Finally, we shift the upper terminals of current generators from the line $+E_c$ to the "grounded" line (Fig. 10.8c). In accordance with the theory of circuits, this shifting will not affect the operation mode of transistors because the current of current generators does not depend on whether the series-connected emf sources (the source E_c here) are present or not.

As seen, a distinguishing feature of I^2L circuits is individual supply of power to the base of each transistor from its "own" current generator. These generators are often omitted for simplicity, and the circuit (Fig. 10.8d) then takes a somewhat strange form as if it were lacking power sources.

Individual current generators use *pnp* transistors connected in the CB configuration (Fig. 10.9a).

This circuit form is known from Subsec. 9.11.2 (see Fig. 9.36b). As shown earlier, this version has its *pnp* transistor **formally** operated in the mode of double injection (saturation) since the collector potential is above zero. But if the voltage V_c is at least 0.1 V below V_e , injection at the collector junction does not essentially occur and the

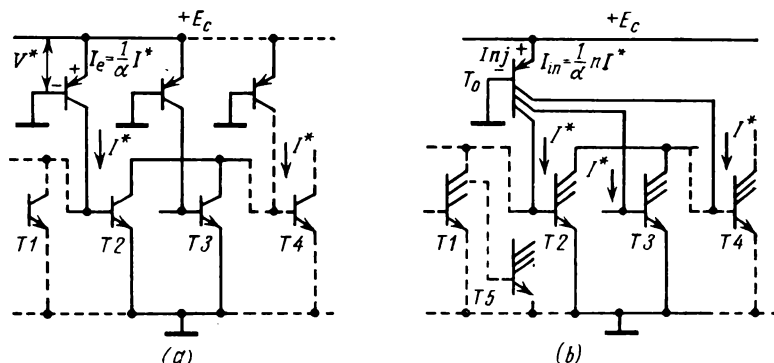


Fig. 10.9. Current generators using individual *pnp* transistors (a) and one multicollector *pnp* transistor (b)

collector current remains constant, equal to αI_e . What ensures a smaller value of V_c over V_e is a noticeable difference between the emitter and collector currents in *pnp* transistors (see below).

As seen from Fig. 10.9a, the emitters and bases of all *pnp* transistors turn out to be connected to each other. The possibility thus exists for replacing *all pnp* transistors by *one np multicollector transistor* (Fig. 10.9b).

Multicollector npn logic transistors are likewise popular in I^2L . The use of these transistors permits eliminating the main drawback of DCTL, namely, the effect of spread in input I - V characteristics. Indeed, the bases of $T2$ and $T5$ in Fig. 10.9b are fed from **different** collectors of $T1$. The bases are thus "isolated" from each other, and so the nonuniform distribution of current between them is unlikely. As a result, the use of multicollector logic transistors greatly facilitates the implementation of intricate connections between the logic elements in complex digital ICs.

The term injection-type supply is quite acceptable for the given IC logic element because the currents I^* result from the **injection** of holes through the emitter junction of *pnp* transistor. The emitter that acts as a current source is customarily called an *injector* and denoted I_{nj} in Fig. 10.9b. Since the voltage E_c is directly fed to the injector junction, then $E_c = V^*$. A low supply voltage is one of the

important advantages of I²L circuits. It is easy to see that this advantage is the consequence of replacement of resistors by current generators.

As known, applying voltage directly to a *pn* junction is objectionable, otherwise it would be difficult to secure a fixed value of current (see p. 89). A practical circuit, therefore, has a low resistance placed in series with the injector. The supply voltage then is little higher than V^* , typically 1 to 1.5 V.

The injector current is distributed among all collectors of the *pnp* transistors, the number n of which can be as large as 10 to 20 and more. So, if the overall gain of injector current is near unity ($\alpha \geq 0.9$), the gain of each collector is $1/n$ as large. This suggests that the currents I^* and I_{inj} differ heavily:

$$I^* \approx (1/n) I_{inj}$$

Consequently, the forward voltage V^* at the emitter junction of an *npn* transistor (that draws current I^*) is lower than V^* at the injector junction. It is exactly this circumstance that enables the *pnp* transistor to meet the condition $V_e - V_c > 0.1$ V at which current stabilization becomes possible despite forward biasing of both junctions (see above).

The logical analysis of I²L circuits commonly involves the use of an equivalent circuit shown in Fig. 10.8c. It is assumed here that the transistor base is either bypassed to "ground" (if in the preceding logic element one of the transistors is biased to saturation) or "open-circuited" (if in the preceding logic element all transistors are held cut off). In the first case, the given transistor is off and the current I^* goes through the transistors of the preceding logic element. In the second case, the current I^* fully flows into the base of the given transistor and drives it to saturation. Let, for example, the transistor in Fig. 10.9b be in saturation. Its base then draws current I^* from the lower collector of T_0 (because $T1$ is cut off) and its collector receives I^* from the upper collector of T_0 (because $T4$ is also off: its base is at a small residual voltage of the saturated transistor $T2$). Consequently, $I_{b2} = I_{c2} = I^*$. The saturation condition (8.4) here takes the form $BI^* \geq I^*$ or

$$B \geq 1 \quad (10.15)$$

As is clear, *the requirements on the base current gain are minimum*. The common condition (10.15) is fulfilled over the microampere region, at currents from 5 to 10 μ A. This is still another important merit of I²L circuit, which, along with a low supply voltage, aids in a sharp decrease of the power consumed.

The logic levels and logic swing in I²L circuits are described by the same formula and have approximately the same values as for the DCTL circuit [see Eqs. (10.7) and (10.8)].

The unique circuit design of I²L is combined with the unique fabrication approach. In Fig. 10.10 are shown the structure and geometry of a typical I²L circuit. For the illustrative purpose, the transistor numbers on the right side of the figure and the connections between transistors correspond to those in Fig. 10.9b.

The role of the emitter, common to all *nnp* transistors, is played by an epi-*n* layer with an *n*⁺ substrate (the latter essentially forms an ohmic contact to the *n* layer). In I²L circuits, there is generally

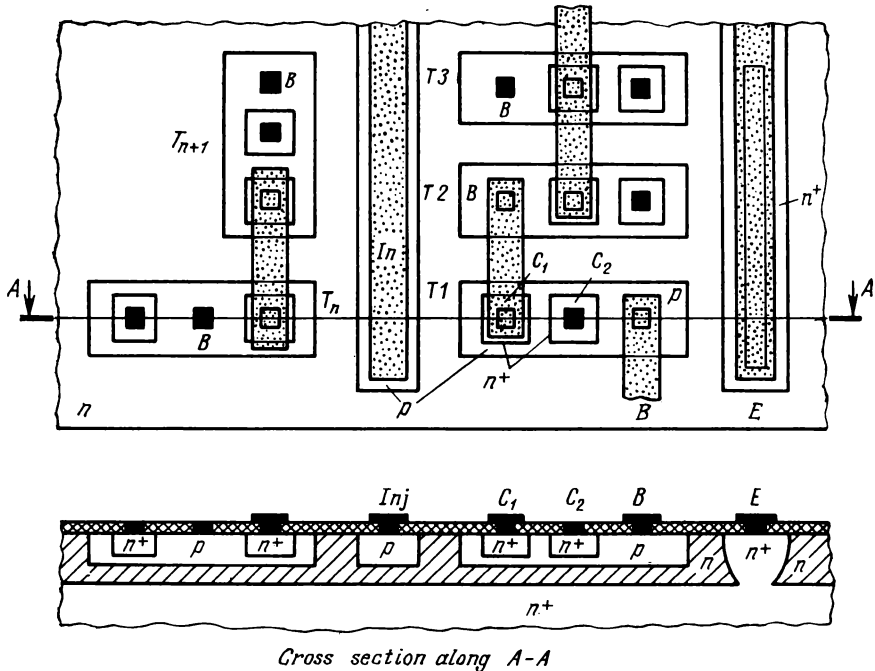


Fig. 10.10. Structure and geometry of an I²L circuit

no need to isolate *nnp* transistors from each other because the common emitter layer is not objectionable but rather necessary as regards the structure of the circuit (see Figs. 10.8 and 10.9).

The injector is fabricated in the form of a long *p* stripe grown at the stage of base diffusion. The *pnp* transistor base is an epi-*n* layer, and the collectors are the base *p* layers of an *nnp* transistor. So, the *pnp* transistor has a lateral structure (see Fig. 7.21) and a homogeneous base, and thus represents a diffusion transistor.

The *nnp* transistors are inverse multiemitter types, the detailed discussion of which was made in Subsec. 7.4.2 (see Fig. 7.17). To raise the normal current gain of a multicollector *nnp* transistor, it

is desirable that the n layer can be made as thin as possible and the p base width be smaller.

The nnp transistor can lie both perpendicular to the injector ($T1$ - $T3$ and T_n in Fig. 10.10) and parallel to it (T_{n+1}). The injector itself should not necessarily be made in the form of a special layer. The p^+ substrate shown in Fig. 10.11 may perform the function of the injector. In this version, the pnp transistor is not lateral as it is

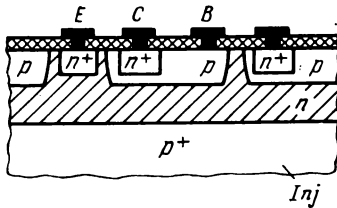


Fig. 10.11. Structure of an I^2L circuit with the substrate acting as an injector

in Fig. 10.10, but **vertical**. A major advantage of this version is a free space available on the chip surface because the injector strip is absent. There are many other versions of the I^2L circuit as regards its design, structure, and geometry.

In summary, the advantages of I^2L are as follows: the absence of insulating islands (space saving); absence of resistors (and hence space saving, a decrease in supply voltage, power dissipation, and delay times); small collector capacitance (due to small areas of n^+ layers); and low residual voltage in the saturated state. The last advantage is due, first, to the fact that the n^+ collector layer is low-resistant [that is, the resistance r_{sc} is low, see Eq. (8.1)] and, second, to the fact that the inverse gain B_I ranges from 100 to 150 and over. Consequently, the summand V_{ce} in (8.1) can be equal to fractions of a millivolt as follows from (4.39).

10.4. MOS Logic

MOS logic circuits presently use MOSFET switches, or inverters, considered in Sec. 8.7.

10.4.1. MOS logic using switches of the same conductivity type. As in Subsec. 8.7.2, we shall focus here on induced n -channel transistors since they operate at positive supply voltage and thus are more convenient for analysis.

Integrated circuit MOS logic elements are easier to analyze than bipolar logic elements because input (gate) circuits of the former do not practically draw current.

So, while operating in the network, *individual logic elements function independent of one another*, and each can be analyzed disregarding the effect of the preceding and the next logic element. In parti-

cular, logic levels V^0 and V^1 do not depend on load and remain the same as they are in the open-circuit condition. The effect of the succeeding (driven) logic element results only in an increase of the output capacitance of the given logic element.

In Fig. 10.12 are shown two standard versions of logic elements employing n -channel MOSTs. Both circuit versions utilize dynamic load because the use of load resistors leads to a sharp increase of the chip area and does away with one of the basic merits of MOS logic—a high level of integration. The logic element of Fig. 10.12a is built

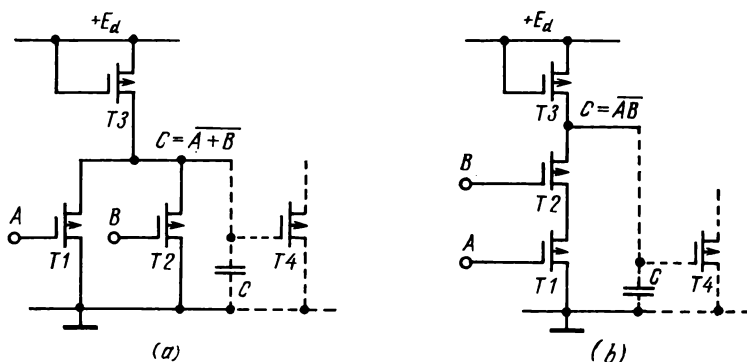


Fig. 10.12. MOS logic elements using switches of the same conductivity type with parallel (a) and series (b) connection of logic transistors

on the same principle as a DCTL element (see Fig. 10.2a): logic transistors T_1 and T_2 are connected in parallel, and when each is turned on, the output level is decreased. The circuit thus performs the NOR function.

The logic levels in a MOS logic circuit represent the output voltages of the switch in the on and off conditions (see Fig. 8.17b). In the on state, the residual voltage at the switch is defined by Eq. (8.49). Given the proper geometry of transistors, this voltage is as low as in bipolar switches (0.05 to 0.15 V). It may thus be assumed that

$$V^0 = V_{res} \approx 0.1 \text{ V} \quad (10.16a)$$

It will be recalled that a small residual voltage presupposes the smallest possible channel width of a load transistor as against the channel width of an active transistor (see p. 294). It should also be noted that in a MOS logic circuit the residual voltage decays in proportion to the number of forward-biased logic transistors, because the parallel connection of transistors is equivalent to an increase in the specific transconductance b_1 in Eq. (8.49).

With the switch off, the output voltage approaches the supply voltage:

$$V^1 \approx E_d \quad (10.16b)$$

Consequently, the logic swing

$$V_l = E_d - V_{res} \approx E_d \quad (10.17)$$

In MOS logic circuits, the supply voltage is usually taken to be three to four times the threshold voltage. So, if $V_0 = 1.5$ to 3 V, then the logic swing (5 to 10 V) far exceeds the swing values specific to DCTL, I²L, ECL, and even DTL and TTL circuits (at supply voltages of 4 or 5 V).

One more advantage of MOS logic is improved noise immunity. Really, the input of an off transistor is at the logic level V^0 . So, to drive the transistor on requires the voltage $V_0 - V^0$ that is near the threshold voltage (1.5 to 3 V), whereas in bipolar logic gates discussed earlier this voltage is 1 or 2 V* (0.7 to 1.4 V).

The MOS circuit of Fig. 10.12b has its logic elements connected not in parallel but in series. That is why the passage of current in the circuit and hence a low output voltage level V^0 are only possible if all logic transistors (two here) go on. This condition prevails when applying the level V^1 to all logic inputs. It is obvious that the given logic element realizes the NAND function.

The level V^1 in this circuit is the same as in the preceding circuit, but the level V^0 is higher: it is proportional to the number of series-connected logic transistors and can be as high as 0.2 to 0.5 V and over. *The logic swing will be smaller* accordingly. At supply voltages of 10 V and above, this drawback is of little significance. But in low-voltage circuits with low threshold voltages, the increased level V^0 presents a certain problem.

As with simple MOS switches, the speed of MOS logic circuits is limited by the recharge rate of output capacitance C (shown by a dash line in Fig. 10.12). The value of C is proportional to the number of driven logic elements and is also dependent on the capacitance of metallization [see Eq. (7.4)].

To raise the switching speed it is required to increase the operating currents of transistors, that is, to increase their specific transconductances. But, as follows from Eq. (5.7), this approach necessitates a larger channel width, that is, a greater area for transistors on the chip. Besides, with the increased operating currents the power dissipation grows, this being one more obstacle on the way to raising the packing density. Because of the difficulties given above, practical MOS logic circuits are not as fast as bipolar gates.

10.4.2. MOS logic using complementary transistors (CMOS logic). Simple CMOS switches were discussed in Subsec. 8.7.3 (see Fig. 8.18).

The basic merit of these devices is that a change in the output voltage is not related to the change in current, which remains near zero. This advantage—very small power demand—is also inherent in CMOS logic elements. Two standard circuit forms of these elements are given in Fig. 10.13. By the principle of action, they are similar to the circuits of Fig. 10.12.

Figure 10.13 clearly shows a regular feature of CMOS circuits: *parallel-connected transistors of one type are followed by series-connected*

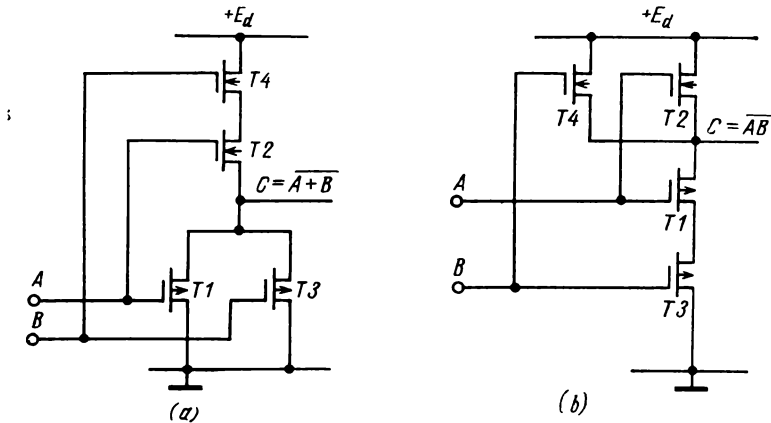


Fig. 10.13. CMOS logic elements using complementary switches with parallel (a) and series (b) connection of logic transistors

transistors of the other type. The function performed is determined by the connection of transistors in the “lower row” (cf. Fig. 10.12). In the circuit under study, these are *n*-channel transistors. On reversing the polarity of supply voltage, *p*-channel transistors will appear in the lower row.

Assume both logic inputs in the circuit of Fig. 10.13a are at level $V^0 < V_0$. In this state, the channels in *n*-channel transistors *T1* and *T3* do not appear (transistors are off). The channels in *p*-channel transistors *T2* and *T4* are produced because the potential difference $V^0 - E_d = V_{gs}$ exceeds the threshold voltage in magnitude. But since negligible currents of the off transistors *T1* and *T3* flow through the channels, the voltage drop across them will be negligible too [see Eq. (8.51b)]. It may then be assumed that the output voltage is equal to the supply voltage. This is exactly a 1 output.

$$V^1 = E_d \quad (10.18a)$$

If an input voltage V^1 is applied to the terminal *A*, a channel builds up in *T1*, but the channel in *T2* disappears (the transistor goes off).

A negligible residual current of $T2$ flows through the channel of $T1$ and practically causes a zero voltage drop. In this case it may be assumed that

$$V^0 = 0 \quad (10.18b)$$

The logic swing then is

$$V_l = E_d \quad (10.19)$$

Apart from high economical efficiency, additional advantages of CMOS logic over MOS logic are low operating voltages (up to $2V_0$ and less) and higher speed (see Subsec. 8.7.4 at the end). The circuit of Fig. 10.13b features similar parameters. Its operation is analyzed by reasoning in the same way as that presented above.

10.4.3. Dynamic MOS logic (DMOS). A common feature of any family of MOS logic circuits is dielectric isolation of logic elements from each other in the circuit due to a great value of input resistance on the gate side. This offers the possibility of developing a specific class of logic elements in integrated circuit form—*circuits of the dynamic type*¹. Let us illustrate this possibility by considering an example of the circuit shown in Fig. 10.12a.

Disconnect the gate of a transistor $T3$ from the supply line and apply to $T3$ trigger pulses with an amplitude $V_{clock} = E_d$ (Fig. 10.14a). These pulses are known as *clock* pulses, and the mode of operation of a logic element at clock pulses as the *synchronous* mode. The IC logic elements discussed above were assumed to be operated in the *asynchronous* mode². It is obvious that in the absence of a clock pulse, when $V_{clock} = 0$, the transistor $T3$ is biased off *irrespective of the state of logic transistors*. In this condition, the power is not drawn from the supply source.

Upon arrival of a clock pulse, the gate of a load transistor practically becomes connected to the supply line; the circuit thus assumes the same configuration as that shown in Fig. 10.12a. *So, when a clock pulse occurs, the synchronous logic element functions in the same manner as the asynchronous one*, and either draws or does not draw power depending on the state of logic transistors.

¹ As applied to digital devices, the term "dynamic" reflects the time factor involved in circuit operation. The fact is that dynamic circuits, as shown by the text, use capacitor charges that are retained for a limited time and primarily determined by leakage current. In "static" circuits such a limitation does not exist: they operate on the principle of changes in voltage or current levels.

² The terms "synchronous" and "asynchronous" refer not only to individual logic elements but also to systems made up of these elements, in particular, to computers. In synchronous systems, all elements, units, and blocks are controlled by clock pulses. Thus, the specifics of the synchronous mode, as described in the text, also relate to systems. It stands to reason that bipolar ICs can also be operated in the synchronous mode provided the loads in the logic elements are made controllable.

From what we have said above it follows that the synchronous mode ensures lower power dissipation. The gain in power is determined by the relative pulse duration, which is the ratio of pulse time period T to pulse width, or duration t_{clock} . The larger the ratio T/t_{clock} , the higher the saving in power. But the pulse width is limited by the time taken for the recharge of parasitic capacitance C , and the pulse period T by the desired message transmission speed (operation rate of a logic element).

In devising a DMOS logic element, the simple circuit of Fig. 10.14a should be furnished with a switch $T5$ which isolates the output of

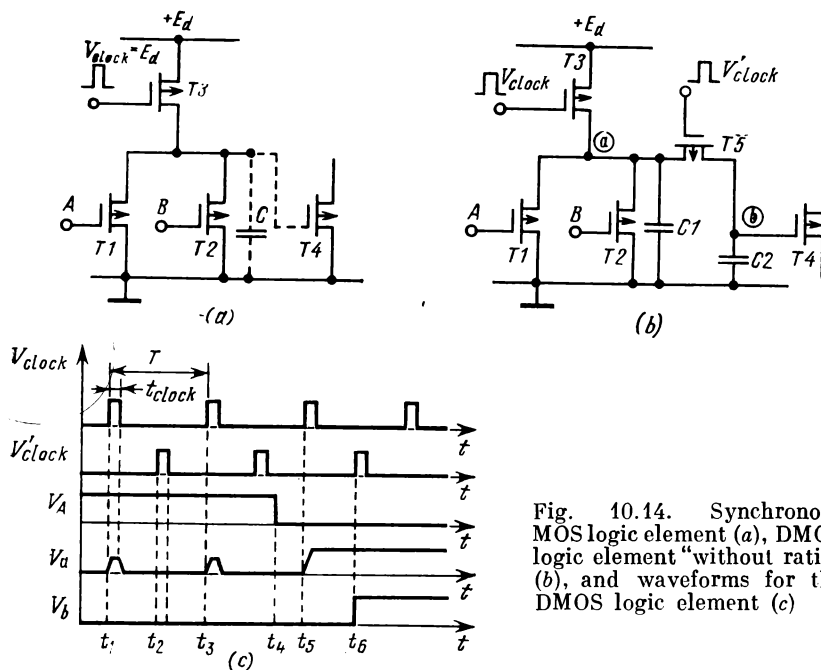


Fig. 10.14. Synchronous MOS logic element (a), DMOS logic element "without ratio" (b), and waveforms for the DMOS logic element (c)

the given logic element from the input of the next one (Fig. 10.14b). The switch $T5$ in combination with capacitors $C1$ and $C2$ forms a memory circuit. The capacitances can be either "parasitic" (which add up to C in Fig. 10.14a) or be specially provided on the chip. The switch $T5$ is controlled by auxiliary clock pulses V'_{clock} shifted relative to the main pulses V_{clock} . The circuit operation is illustrated by pulse waveforms shown in Fig. 10.14c. Assume for simplicity that the input cut-off voltage V^0 at terminal B remains **invariable** and the voltage at terminal A takes a value of V^1 or V^0 . Suppose $V_A = V^1$ in the initial condition (that is, $T1$ is on). Then, in the interval be-

tween clock pulses when $T3$ is biased off, a negligible current flows through $T1$ and the residual voltage at the output (at a) is practically equal to zero [see Eq. (8.51b)].

Upon arrival of the next pulse V_{clock} (at t_1), $T3$ switches on, and a residual voltage sets in at point a , which depends on the ratio between the specific transconductances of an active and a load transistor [see Eq. (8.49)]. If the dimensions of transistors are comparable, this voltage may be rather large. As the pulse V_{clock} ceases, V_a again drops to zero.

When a pulse V'_{clock} occurs (at t_2), the switch $T5$ goes on. In this condition, $C1$ and $C2$ become connected in parallel, with the same voltage being maintained on each capacitor (equal to zero in the given case). After cessation of the pulse V'_{clock} , the switch $T5$ becomes off and $C2$ remains at the zero voltage despite the fact that when the next pulse V_{clock} arrives at moment t_3 , the voltage V_a again rises temporarily.

Thus, at the input of the next logic element (at b) the level V^0 is equal to zero irrespective of the ratio b_1/b_3 for the preceding logic element. This circumstance permits us to choose the ratio b_1/b_3 close to unity, that is, to decrease the dimensions of the active transistor to those of the load transistor and thus save a considerable amount of space on the chip. This is one of the important advantages offered by DMOS logic.

The DMOS logic circuits built according to the described principle are referred to as circuits "without ratio", and the simple circuit of Fig. 10.14a, where the inequality $b_1/b_3 \gg 1$ must be met, is known as a circuit "with ratio".

With the voltage $V_A = V^0$ applied (at t_4) to the terminal A , the transistor $T1$ switches off, and the next pulse V_{clock} arriving at the moment t_5 causes V_a to rise to the level $V^1 = V_d$. This level then keeps constant because $T1$ stays off. When the switch $T5$ assumes the on condition anew (at t_6), the capacitor $C2$ charges to the same voltage level¹, which does not change after driving $T5$ into cutoff.

From the above description we can conclude that in the circuit "without ratio", information is transferred from one logic element to the other with a **delay**, that is, with a shift by one clock period.

10.5. Logic Element Parameters

All logic elements are defined by a certain set of parameters given in reference books and other kinds of documentation. Most of the parameters have a clear, officially adopted, definition, which excludes the

¹ With C_2 in parallel with C_1 , the voltage generally drops off because of charge distribution between the capacitances. But this drop in voltage is insignificant if $C_1 \gg C_2$.

ambiguity of measuring techniques and enables comparison of the different types of logic element. The full list of parameters is far too long. Therefore, we shall introduce the reader to a limited number of parameters to begin with; namely, to the parameters which are most important for comparative estimations.

Mean power dissipation per gate. Its expression has the form

$$P = 1/2 (P^0 + P^1) \quad (10.20)$$

where P^0 and P^1 are the powers consumed by an IC logic element in the logic 0 and logic 1 states respectively.

The definition of the mean power relies on the fact that in a complex multiple-element device, **on the average**, one half of the logic elements assumes the 1 condition and the other half the 0 condition simultaneously. Thus, multiplying the value of P per gate by the number of gates yields the desired source power.

From Eq. (10.20) it is seen that the mean power P is a **static** quantity: it does not include the power lost during short transients when the logic element goes from logic 1 to logic 0 and vice versa.

Mean propagation delay time. This is the switching speed of a gate given by

$$t_{pd} = 1/2 (t_{pd}^{1,0} + t_{pd}^{0,1}) \quad (10.21)$$

where $t_{pd}^{1,0}$ and $t_{pd}^{0,1}$ are the mean propagation delays between the leading edges of output and input waveforms at switch-on (when V_{out} changes from V^0 to V^1) and at switch-off (when V_{out} changes from V^1 to V^0). Delays are measured either at the 50% level of the maximum pulse amplitude or at the level of sensitivity threshold (see p. 300).

Since a transient depends on the nature of a load, the delay time is estimated under definite output conditions assuming that the load of the given logic element is the input circuit of a similar logic element.

Mean power-delay product. This parameter (also known as a speed-power product)

$$A = Pt_{pd} \quad (10.22)$$

characterizes both the economical efficiency and the speed of an IC logic element. At present, it is by this parameter that one makes the first-priority comparison of the different types of logic element and, in particular, the estimation of new types as regards the promise they hold.

The power-delay product as a parameter has a definite physical meaning. The simplest way to reveal this aspect is to consider an example of the DCTL circuit shown in Fig. 10.2a, assuming that the mean power dissipated by this IC gate is

$$P = E^2/R_c$$

Table 10.1

Basic Parameters for Integrated Logic Elements

Logic family	P , mW	t_{pd} , ns	A , pJ	V_{nst} , V	K_{f-in}	K_{f-out}
TTL Schottky TTL	1-20	$\frac{5-20}{2-10}$	$\frac{50-100}{20-50}$	$\frac{0.8-1}{0.5-0.8}$	2-5	10
ECL	20-50	0.7-3	20-50	0.2-0.3	2-5	10-20
I ² L	0.01-0.1	10-100	0.2-2	0.02-0.05	1	3-5
MOS CMOS	$\frac{1-10}{0.01-0.1}$	$\frac{20-200}{50-100}$	$\frac{50-200}{0.5-5.0}$	$\frac{2-3}{1-2}$	2-5	100-200

(where E is the supply voltage) and the mean propagation delay is

$$t_{pd} = R_c C_c$$

where C_c is the overall capacitance connected to collectors¹. Then, disregarding the numerical coefficients (which should have been entered into the expressions for power dissipation and delay time), the mean power-delay product will take the form

$$A \approx E^2 C_c \quad (10.23a)$$

Considering that the capacitance is first of all proportional to the area of a transistor and assuming, for clarity, that the transistor is square in configuration, we rewrite Eq. (10.23a) to obtain

$$A \approx E^2 a^2 \quad (10.23b)$$

where a represents linear dimensions of the transistor.

It is obvious from Eq. (10.23) that the power-delay product describes the *physical-manufacturing and circuit engineering level of integrated circuits* because the area of a device and the operating voltage depend in the final analysis on the device type, its structure, the resolution of photolithography, and other analogous factors.

At the **given** physical-manufacturing and circuit engineering level, with A being specified, it is possible to realize logic elements in various circuit configurations. Unfortunately, as seen from Eq. (10.22), they will show either a *high speed at low efficiency* or, on the contrary, a *high efficiency at low speed*. This inverse relationship is well known to IC development engineers and is clearly evident from Table. 10.1.

¹ This definition of the delay time does not allow for either the storage time or the transit time for carriers in the base, that is, the delay is assumed to be as small as possible.

The principal way of advancement in the physics, technology, and circuit engineering of microelectronics must be the way of decreasing the power-delay product. This statement is fairly evident from the experience gained over the last 10-12 years in raising the packing density of elements (which had led to a reduction in the power-delay from 50-100 pJ to 2-5 pJ), from the development of I²L (see Table 10.1), and from the researches aimed at designing unique electronic devices quite distinct from transistors. For example, the use of the *Josephson effect* permits in principle decreasing the power delay product to 10⁻⁴ pJ and below.

Static noise immunity. This parameter $V_{n\ st}$ is a maximum possible voltage of **static** noise at which changes in the output levels of an IC logic element do not yet take place. By static noise are meant parasitic voltages and currents whose duration is larger than the time it takes to switch the logic element from one stable state to the other. The mechanism of this noise and the methods for its analytical estimation were discussed in Sec. 8.8.

Noise immunity is sometimes evaluated as a ratio of $V_{n\ st}$ to the logic swing. This parameter

$$K_{n\ st} = V_{n\ st}/V_l \quad (10.24)$$

is called the **noise immunity factor**.

Fan-in K_{f-in} . This is the number of inputs that can be connected to a logic gate. There is a certain limit to the number of inputs (for example, to the number of transistors in an RTL gate or emitters in a TTL gate). This limit is due not only to design and manufacturing factors but also due to the interaction between inputs and to increased delays (because each logic input introduces an additional capacitance).

Fan-out K_{f-out} . This is a permissible number of loads that can be driven from one output of the given integrated logic element. The load is an input of a similar logic element. An increase in the fan-out is limited by the operating current, spread in parameters, and also by higher overall output capacitance and thus longer delays.

If there is a need to increase the fan-in and fan-out, special buffer stages, known as *input* and *output expanders*, are connected to logic elements at the inputs and outputs respectively. With these facilities, the gates are said to be expandable.

The typical parameters of basic modern IC gates are listed in Table 10.1. The data illustrate those features of individual logic families which were noted above in respective sections. ECL circuits and Schottky (barrier) TTL circuits are fastest; DMOS circuits and, in particular, CMOS circuits are most economical. In the bipolar logic family, I²L circuits are most economical in operation.

10.6. IC Flip-Flops

In Sec. 8.9 we have shown that flip-flops employ switches with a positive feedback loop placed around them. From Fig. 8.28a it is seen that the RS flip-flop consists of two pairs of transistors. One transistor in the pair is a triggering transistor and the other is included into the feedback loop. Now that we are familiar with logic elements, it can be readily seen that each pair in the RS flip-flop represents

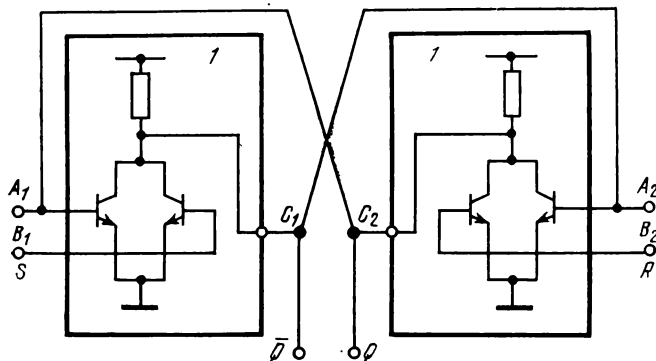


Fig. 10.15. RS flip-flop using two NOR gates

a two-input DCTL NOR gate (the NOR gate using transistor logic in DCTL form). For clarity, Fig. 10.15 shows the same flip-flop with the use of the standard designations for NOR logic circuits (see Fig. 10.1d).

The conclusion we have just made is general in character: *any flip-flop is a combination of a few logic elements connected in a definite manner*. The number of IC logic elements used and the methods for their connection differ with each type of flip-flop. Correspondingly, the functions they perform differ too. Along with RS flip-flops, therefore, there is a rather large variety of other flip-flops. The type of logic element used in a flip-flop determines such basic parameters as the switching speed, power dissipation, loading factor, and others.

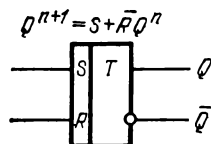
In the above sections we have treated in detail the schematic diagrams, parameters, and features of logic elements. Therefore, here we shall restrict ourselves chiefly to **block** diagrams in order not to distract the attention and look into the details of schematic diagrams, the more so as the latter have a rather complex configuration if they involve a large number of gates.

10.6.1. RS flip-flop. The logic circuit configuration of an RS flip-flop using NOR gates is shown in Fig. 10.15. Its logic formula and block symbol appear in Fig. 10.16. The superscripts n and $n + 1$

stand for the values of the Q output before and after arrival of clock signals, that is, during the n th and $(n + 1)$ th clock period.

As known, the voltage levels at both flip-flop outputs are different and **simultaneously** change to logically opposite levels. That is why in the flip-flop symbol, one of the output is labelled Q and the other \bar{Q} (the condition of inversion is generally denoted by a circle on the side of a rectangle). The Q output is regarded to be the

Fig. 10.16. Symbol for an RS flip-flop



main output: the values of Q represent the state of the trigger as a whole. Thus the statement “the trigger assumes the 1 condition” means that the Q output = logic 1 (\bar{Q} = logic 0).

In order to avoid repetition in the further discussion, let us verify the logic formula (see Fig. 10.16):

$$\begin{aligned} \text{if } S = 0, \quad R = 0, \text{ then } Q^{n+1} &= 0 + 1 \cdot Q^n = Q^n \\ \text{if } S = 0, \quad R = 1, \text{ then } Q^{n+1} &= 0 + 0 \cdot Q^n = 0 \\ \text{if } S = 1, \quad R = 0, \text{ then } Q^{n+1} &= 1 + 1 \cdot Q^n = 1 + Q^n = 1 \end{aligned}$$

One more possible set of conditions ($S = 1, R = 1$) will be considered a little later.

Based on the conditions given above, we arrive at the following conclusion: the signals to both S and R inputs (signals represent logic 1) *provide an unambiguous state of the flip-flop*. The signal $S = 1$ causes $Q = 1$ and signal $R = 1$ causes $Q = 0$. As the signal ceases, the accepted state does not change.

Turn now to the set of conditions $S = 1$ and $R = 1$. Whatever the preceding state of the flip-flop can be, these input signals, as clear from Fig. 10.15, cause **the same** logic levels at the outputs: $Q = \bar{Q} = 0$. This circumstance alone points to an abnormal situation. But the main contradiction lies in the fact that *after cessation* of the S and R signals, the flip-flop assumes an *indefinite* state: at the first moment both pairs of inputs stay at the logic 0 levels. Being affected by internal fluctuations, the flip-flop may pass to any of the two stable states with equal probability: $Q = 1$ or $Q = 0$. This fact was duly discussed in Subsec. 8.9.2. For this reason, the set $S = 1, R = 1$ in the given flip-flop is *inhibited*, that is, it must not be encountered in the flip-flops used in electronic equipment. This inhibition is clear from general considerations: *it is prohibited to*

apply simultaneously opposite instructions to the flip-flop: "set to 1 (S)" and "reset to 0 (R)".

RS flip-flop can certainly use not only NOR gate circuits but also NAND gate circuits (for example, in TTL form). NOR gate circuits can evolve into NAND gates by changing all the input and output variables to logically inverted variables (see p. 378). The RS flip-flop composed of NAND gates will have a configuration as shown in Fig. 10.17a. The logic formula for this flip-flop circuit form is the

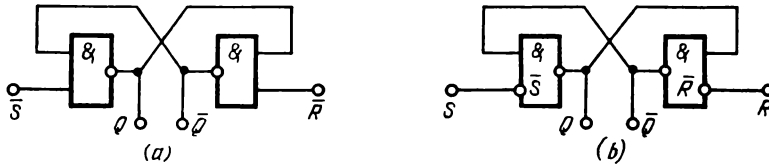


Fig. 10.17. RS flip-flop using two NAND gates

(a) circuit evolved by inverting input and output quantities of circuit in Fig. 10.15; (b) same circuit with internal input inverters

same as that given in Fig. 10.16. But in contrast to the circuit of Fig. 10.15, here *the position of the main output Q has changed*, and the inputs have to receive *inverse signals \bar{S} and \bar{R}* . If the inverter forms part of an integrated logic element circuit, the symbol of the latter changes—a circle appears at the input (Fig. 10.17b).

Schematics of RS flip-flops made in I^2L and T^2L forms are given in Fig. 10.18: The first circuit (see Fig. 10.18a) employs NOR logic

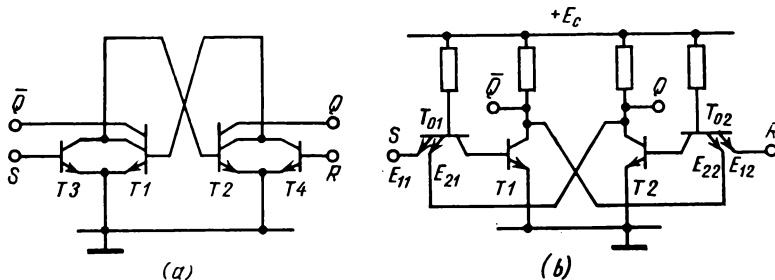


Fig. 10.18. I^2L RS flip-flop (a) and TTL RS flip-flop (b)

elements and operates in the positive logic condition; the second circuit (Fig. 10.18b) is a negative logic NAND-gate flip-flop.

10.6.2. RST flip-flop. The RS flip-flop considered above belongs to the class of *asynchronous* circuits which change state only with a change in the level at the appropriate input. In wide use are *synchronous flip-flops*, or RST flip-flops, which can change state only

after arrival of special, *clock pulses* (for comparison see Subsec. 10.4.3). Within the clock-pulse space width, changes in the levels at the *S* and *R* inputs do not cause changes in the condition of the flip-flop, but only “program” that state which the circuit will assume on applying the next clock pulse.

In the notation of RST flip-flops, the letter *T* denotes the toggle input that receives clock pulses (in the logic diagrams, therefore, the

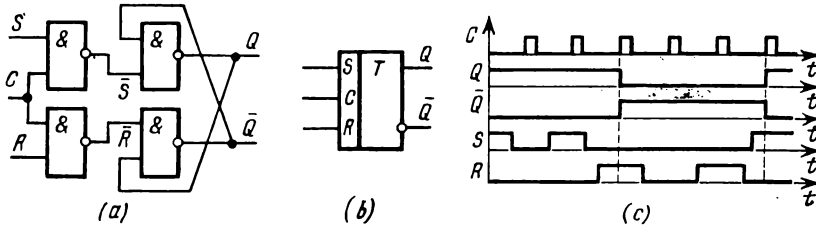


Fig. 10.19. RST flip-flop
(a) block diagram; (b) symbol; (c) waveforms

T input is identified as *C* for clock). The block diagram, logic symbol, and operating waveforms of an RST flip-flop are presented in Fig. 10.19.

As is clear, the circuit is basically a set-reset arrangement made up of NAND gates (see Fig. 10.17a). The inputs of this flip-flop are controlled by two more NAND gate circuits which invert the *S* and *R* levels at each clock pulse *C*. If a clock pulse does not occur ($C = 0$), the NAND circuits remain inactive and the RST flip-flop does not change state. The effect of a clock pulse may be defined by a logic formula

$$Q^{n+1} = C(S + \bar{R}Q^n) \quad (10.25)$$

10.6.3. T Flip-flop. In Subsec. 8.9.3 we have mentioned that a common-input flip-flop is only possible if the circuit has an internal memory. Fig. 8.30 shows such an internal memory provided by two memory capacitors. As known, it is usual practice to strive to avoid the use of capacitors in integrated circuits. Thus it is necessary to secure the memory by a circuit design approach.

The block diagram of an IC flip-flop with a common input (T flip-flop), and also its logic symbol and waveform are illustrated in Fig. 10.20. It is obvious from the figure that the T flip-flop consists of two RST flip-flops (*M* and *N*) and an inverter which feeds an inverted clock pulse to the *N* flip-flop. In essence, this is a master-slave flip-flop, *M* being a *master*, and *N* a *slave*.

In the interval between clock pulses (when $C = 0$), the output levels in both flip-flops are equal: $Q = Q_1$. Assume, for example,

that $Q = Q_1 = 1$ (see the initial state in Fig. 10.20c). So that the next incoming clock pulse will change the state of the master flip-flop (that is, provide $Q_1 = 0$), it is first necessary to establish corresponding levels at its inputs: $S_1 = 0$ and $R_1 = 1$. This function is performed by cross feedback paths going from the output of the slave to the input of the master. Indeed, from Fig. 10.20a it follows: $S_1 = \bar{Q} = 0$ and $R_1 = Q = 1$. Thus, *in the interval between clock pulses the master is ready to toggle as the next clock pulse occurs (at t_1 in Fig. 10.20c).*

As for the slave flip-flop, its state cannot change at the time of clock pulse since during this time its clock input is disabled by an

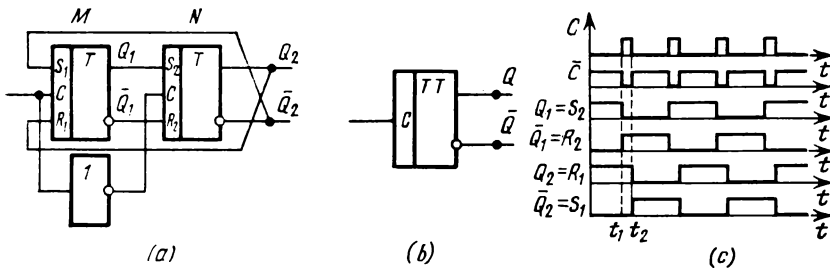


Fig. 10.20. T flip-flop

(a) block diagram; (b) symbol; (c) waveforms

inhibit signal $\bar{C} = 0$. Consequently, at the time of the clock pulse the Q and \bar{Q} outputs do not change and, hence, nor do the S_1 and R_1 inputs. The last circumstance ensures reliable locking of the master in either of the states.

Figure 10.20a clearly indicates that switching of the master entails a change in the S_2 and R_2 levels at the slave inputs. That is why *by the end of the clock pulse propagation, the slave becomes prepared to switch to a new condition that corresponds to the new condition of the master.* Such switching occurs when the clock pulse ceases and an enabling signal $C = 1$ (at t_2 in Fig. 10.20c) appears at the clock input of the slave.

Thus, *every clock pulse causes the T flip-flop to assume a new stable state but with a delay equal to the clock pulse duration (the same shift occurs in the classical T flip-flop with a capacitive memory, see Fig. 8.30b).* The logic formula for the T flip-flop may be written as

$$Q^{n+1} = CQ^n + \bar{C}\bar{Q}^n \quad (10.26)$$

where CQ^n is the value at the time of the clock pulse; and $\bar{C}\bar{Q}^n$ is the value after the clock pulse cessation.

10.6.4. JK flip-flop. The jump-keep (JK) flip-flop is most versatile. The J and K inputs, like S and R inputs, provide the desired state, but, in distinction to an RS flip-flop, this circuit allows for the set of conditions $J = 1, K = 1$. By the principle of action, JK flip-flops belong to the category of **synchronous** devices: the output sevels set in only after arrival of clock pulses. The block diagram, lymbol, and waveforms of a JK flip-flop are illustrated in Fig. 10.21.

The JK flip-flop is seen to be built from a master-slave T flip-flop. However, the S and R inputs here are cross-connected to the outputs

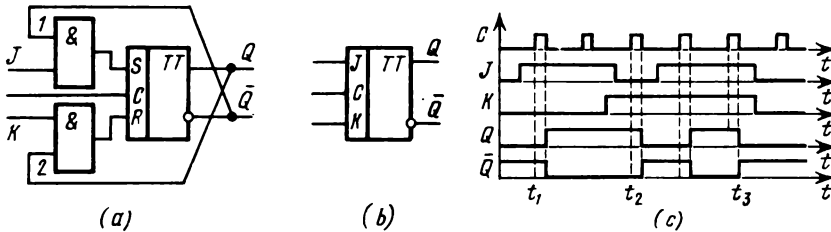


Fig. 10.21. JK flip-flop

(a) block diagram; (b) symbol; (c) waveforms

not directly, as the T flip-flop, but via two-input AND circuits, with one input of each intended to receive signals J or K . With input levels 1 at both J and K , the AND circuits convert to followers transmitting the levels applied to 1 and 2, and the unit behaves like a T flip-flop (t_2 - t_3 in Fig. 10.21c). With different input levels, the circuit behaves like an RST flip-flop (interval t_1 - t_2) where the level J provides $Q = 1$ and the level K provides $Q = 0$.

The logic formula for a JK flip-flop is

$$Q^{n+1} = J\bar{Q}^n + \bar{K}Q^n \quad (10.27)$$

In particular, at $J = 1, K = 1$, we have $Q^{n+1} = \bar{Q}^n$ [the same as in the T flip-flop at $C = 1$, see Eq. (10.26)].

10.6.5. D flip-flop. The characteristic feature of a delay (D) flip-flop lies in that it does not change state until the arrival of the next clock pulse. Fig. 10.22 shows the block diagram of a D flip-flop, along with its logic symbol and waveforms.

Obviously, the D flip-flop is a JK circuit modification with the J input connected to the K input via an inverter. The K input is thus **dependent**: a signal to this input is related to a signal at the main input by $K = \bar{J} = \bar{D}$. In the interval between the clock pulses when $C = 0$, the signal at the input D can either remain

unchanged or change to an opposite signal. In both cases, the Q and \bar{Q} levels do not alter until the next clock pulse arrives. The next clock pulse ($C = 1$) causes the output levels to change (or not to change) according to the existing level $J = D$ (since the level K is unambiguously determined by the level J).

If at the moment of arrival of the $(n + 1)$ th clock pulse $D = 1$, then $J = 1$ too, and hence $Q = 1$ in accordance with the principle of the JK flip-flop (see Subsec. 10.6.4). If at the moment when the

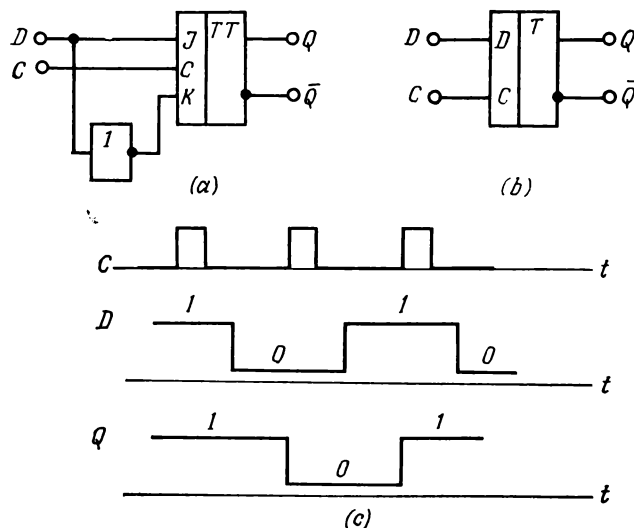


Fig. 10.22. D flip-flop

(a) block diagram; (b) symbol; (c) waveforms

$(n + 1)$ th clock pulse occurs $D = 0$, then $J = 0$, $K = 1$ and $Q = 0$. From the above reasoning it follows that the logic formula for a D flip-flop has a simple form

$$Q^{n+1} = D^n \quad (10.28)$$

The discussed version of a D flip-flop is known as a *D latch*. What distinguishes this circuit is that a change in the D level *at the time* of the clock pulse changes the output levels. In a more complex version of the D flip-flop, the output levels are determined by that input level which exists *at the beginning* of the clock pulse (on its *leading edge*); a change in the D level at the time of the clock pulse does not affect output levels.

10.7. Memories

In digital devices (primarily in computers) memory systems, or memories, hold a most important place. Memories can be *external* and *internal*. The storage media used so far in external memory are magnetic tape and magnetic disks or drums. Internal memory structurally forms an integral part of electronic units. Until recently, it employed ferrite cores and presently has converted to transistor bistable units as a storage medium. External magnetic memories can store information for an indefinitely long period of time and can practically have an unlimited *storage capacity* in terms of bits. A bit is abbreviation for binary digit, which is a unit of information equal to two possible values or states 0 or 1 in the binary number system.

Internal memories are mainly intended to store intermediate data in the process of handling arithmetic or logical problems and also store small standard programs necessary for a given digital device when solving **typical** problems. The former are known as *random access* memories (RAMs) and the latter as *read-only* memories (ROMs).

Random access memories enable a fast alternate data input and output (writing and readout); they have any individual memory (storage) cell equally accessible both for writing and readout.

Unlike RAMS, *ROMs* are basically used for retrieval of stored information. Writing is made either "once and for all" or, in any case, very rarely.

10.7.1. Random access memories. Any RAM comprises two parts, a *storage* and *control units* known as *peripherals*. The storage, used primarily for storing information in binary codes, is a basic and specific section of the RAM. The peripherals are devices for setting and retrieving the data. They include decoders, amplifiers, registers, various keys, commutators, and other general-purpose units. We shall not consider peripherals here and center our attention on the storage.

The storage consists of *memory cells* (MCs) each holding one bit of information, as a binary 0 or 1. Naturally, it is bistable units that form the basis of a memory cell since they feature two stable states, $Q = 1$ or $Q = 0$.

Figure 10.23 shows the typical *matrix* circuit of RAM where every individual cell is located in the nodes of a "network" formed by *address lines* or *wires* X and Y . The number of cells is equal to the product of the number of horizontal by the number of vertical lines (for example, $4 \times 4 = 16$ cells). Each memory cell is connected to one horizontal and to one vertical address line. Hence, if voltages are applied to a definite pair of lines (to X_1 and Y_2 , for example), then quite a definite cell becomes connected to peripheral circuits (in the example considered, this is the MC_{12} shown as a hatched square in Fig. 10.23). A required bit of data (0 or 1) can be inserted into this cell (with

a **unique** address X_1Y_2) or retrieved from it (the cell MC_{12} holds a binary 1).

Both writing and readout are carried out using *bit lines* or wires BL_1 and BL_0 which are connected to all MCs¹. The subscripts attached to the symbols of bit lines are conventional to a certain extent: it is quite possible to apply voltage levels V^0 or V^1 to either of the lines, the binary-1 line BL_1 or binary-0 line BL_0 . Thus the index numbers do not predetermine at all the logic level at any line but conditionally point to the fact that one of the lines (BL_1) is connected to

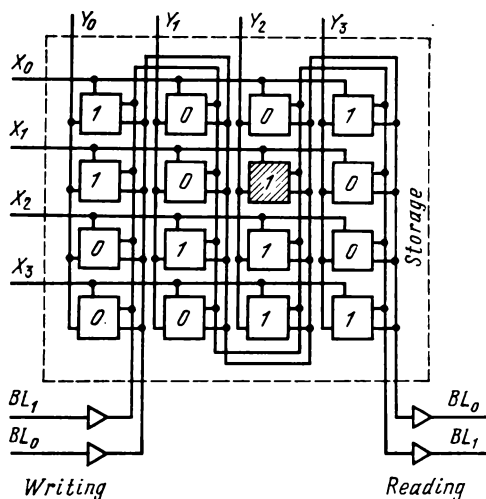


Fig. 10.23. RAM with matrix interconnection

flip-flop **main** outputs Q , while the other (BL_0) to their counterparts \bar{Q} . In writing, the readout outputs are disconnected and, after addressing, one of the two possible codes (01 or 10) is applied to bit lines depending on what information (0 or 1) is to be stored in the selected MC (in Fig. 10.23, the digits in the squares stand for the levels at the cell **main** outputs). In readout, the data inputs are disconnected and so after addressing the levels stored in the selected memory cell are transferred via amplifiers to requisite peripheral units.

There is a great variety of memory cells that make up the storage section of a RAM. Examples of such cells appear in Fig. 10.24.

The cell using single-type p -MOS transistors (Fig. 10.24a) features a classical RS flip-flop structure with control switches $T5$ and $T6$. These switches are normally reverse biased, and the cell is discon-

¹ At the inputs and outputs of bit lines, the triangles identify commutators which connect these lines to (or disconnect them from) control devices (see below).

nected from bit lines. With a negative-going pulse $-E_d$ applied to the address line, $T5$ and $T6$ turn on and connect the cell to bit lines. As this takes place, the levels Q and \bar{Q} written in the cell transfer to the bit lines. In the data writing mode, the address line also receives a pulse $-E_d$, but now the bit lines are given the required (mutually opposite) logic levels which drive the cell into a definite state. Thus in both modes of operation, a pulse in the address line plays the role of a clock pulse.

Figure 10.24b shows a **dynamic** RAM cell, in which a bit of information is stored by means of capacitances C_1 and C_2 (these are generally parasitic capacitances in MOS transistors). The techniques of

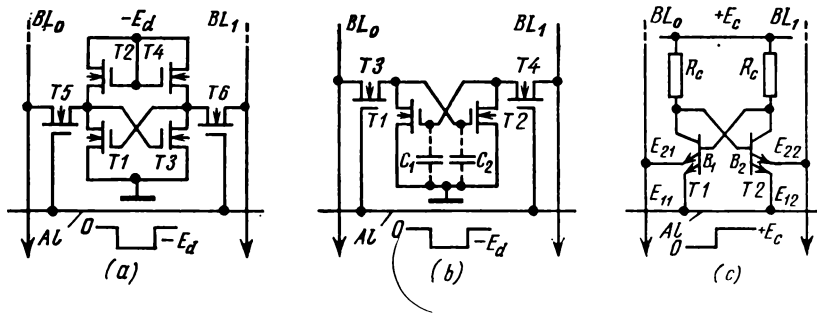


Fig. 10.24. RAM memory cells

(a) static, using single-type MOS transistors; (b) dynamic, using single-type MOS transistors; (c) static, using multiemitter bipolar transistors

writing and readout are the same as in the static cell. Assume the levels $-E_d$ and 0 are applied to lines BL_1 and BL_0 , respectively, to write information into the cell. The pulse of level $-E_d$ will move via the switch $T4$ to the gate of $T1$ causing it to turn on. The gate of $T2$ will receive a pulse of level 0, and the transistor will go off. The voltages across the capacitors will be equal to $V_{d1} = -E_d$, $V_{d2} = 0$.

If the residual current in the off transistor $T2$ is sufficiently small, C_1 will discharge rather slowly and, hence, the voltages $-E_d$ and 0 will be retained for a long time at the cell outputs (at the drains). This time period is sufficient to refresh the data a few times (though during readout the capacitance is additionally shunted by read circuits and its discharge is accelerated). In order to maintain the voltage across the capacitor despite its inevitable discharging, *regeneration* is necessary, that is, restoration of information by periodically writing the data of the same code. Dynamic cells are much more economical than static ones because they do not contain a supply source and thus do not draw power in the data storage period.

On the whole, MOS cells are more economical and compact than bipolar counterparts, but inferior to the latter in speed. Therefore,

though MOS memories are more popular, bipolar memories hold an important place too. An example of the cell using multiemitter bipolar transistors is given in Fig. 10.24c. The principles of writing and readout here are the same as in the above described memory cell, excepting the positive polarity of logic levels and of an address pulse.

Suppose in the data storage mode the transistor $T2$ is off and $T1$ in saturation, so that $V_{b1} = V^*$ and $V_{b2} \approx 0$. If bit lines are at a small "guard" potential (0.1 or 0.2 V), the emitter junction E_{21} will be practically reverse biased and all the current will flow through E_{11} ; in $T2$ both emitter junctions will be in the off condition.

In the readout cycle (with the positive voltage E_c applied to the address line), the emitter E_{11} turns off and the current of $T1$ goes into the line BL_0 via E_{21} which remains at a low potential; the line BL_1 remains dead.

In writing, along with an address pulse, a voltage $+E_c$ is simultaneously applied to that bit line which is connected to a transistor subject to reverse biasing. In the given example, if it is the line BL_1 that accepts the voltage $+E_c$, the transistor $T2$ stays off and the cell will not change state. But if the voltage $+E_c$ is impressed on the line BL_0 , both emitters of $T1$ will be reverse biased. The current then flows via the base of $T2$ into E_{22} , which is at a low potential of the line BL_1 . The transistor $T2$ now switches on, that is, the cell assumes the opposite state.

The list of parameters for integrated RAMS includes the following basic quantities.

Information capacity in terms of bits. This parameter characterizes the level of integration of elements on the chip.

Power per bit. This is the total power consumed in the storage mode, as referred to 1 bit.

Minimum access time T_{ac} . This is a minimum period from the beginning of one read cycle to the beginning of the next. The quantity that is the inverse of T_{ac} is known as the *access rate*. In writing, both these parameters may be somewhat different.

Relative cost per bit of information. This is the total chip cost divided by the information capacity. This parameter is a deciding one in comparative estimations.

Table 10.2 lists the above parameters for several typical families of integrated RAMs.

Comparing the given parameters permits us to make the following conclusions.

MOS RAMs generally excel bipolar counterparts in information capacity, power per bit, and cost per bit, but are much inferior in speed. Among bipolar RAMs, I²L circuits hold a special place. They are next to MOS memories in storage capacity and power per bit. Among the latter, CMOS circuits exhibit a minimum specific

Table 10.2

Basic Parameters for Integrated RAMS

Parameter	MOS transistor			Bipolar transistor		
	static	dynamic	CMOS	TTL	ECL	I ² L
Capacity, bits/IC	1 024-4 096	(4-64) 10 ³	1 024-4 096	1 024-4 096	1 024-4 096	(4-16) 10 ³
Power per bit, mW/bit	0.5-1	0.02-0.1	0.001-0.01	0.3-0.8	0.5-2	0.05
Access time, ns	100-500	100-1 000	50-500	10-20	5-10	200
Cost per bit, relative units	5-10	1-2	10-50	20-50	50-100	10

power, and dynamic MOS RAMs a minimum cost per bit. Of bipolar memory types, ECL memories have the highest speed.

10.7.2. Read-only memories (ROMs). As mentioned earlier, in this type of memory the writing of information is made either "once and for all" or represents a special, rarely performed operation. So, this storage arrangement is primarily employed for information-retrieval applications.

The typical diagram of a diode ROM shown in Fig. 10.25 is of the matrix type where address lines form the rows and bit lines

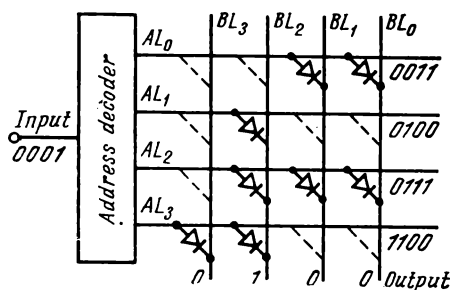


Fig. 10.25. Diode-type ROM

the columns. Every address line stores a definite code: 0011, 0100, etc., as shown in the figure. Code writing is performed with the aid of diodes connected between address lines and between those bit lines which must store logic 1 in readout.

Let the address decoder have selected an address line AL_1 (the plus sign in Fig. 10.25). The voltage in this line is then fed to bit line BL_2 ; the voltage at BL_0 , BL_1 , and BL_3 will be zero. Hence, in parallel reading of information from all the four bit lines we obtain the code (word) 0100 written in the chosen row.

In designing integrated matrices of ROMs it is inexpedient to position diodes just in the nodes where they implement the desired codes. The range of matrices with various versions of code sets is too large, while the run of each version is too small, unjustifiable from the viewpoint of economy. For this reason, diodes are disposed at **all** nodes of the matrix, and in such a **homogeneous** form the matrix is delivered to the customer. The customer himself writes the desired codes into the ROM. For this (using special facilities), he *burns out* the output leads (bridges) of those diodes which are located at the sites of logic 0. The diodes with burnt leads are shown

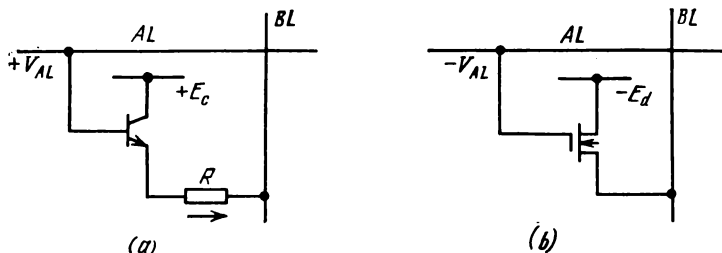


Fig. 10.26. Transistor-type ROM memory cells using a bipolar transistor (a) and MOS transistor (b)

in Fig. 10.25 by dash lines. The leads are burned individually by passing through the appropriate diodes a heavy current that exceeds the normal current rating. To prevent the portions of address and bit lines adjacent to diode **leads** from burning, the diode leads are made higher-resistant and easier-melting than the lines.

Though they are simple in structure, ROMs suffer from a disadvantage that the desired current in bit lines must be provided by a decoder which transfers this current via an address line. To facilitate the decoder work, diodes are replaced by amplifying devices, transistors. Two typical examples of memory cells are given in Fig. 10.26. In using bipolar transistors (Fig. 10.26a), an address line carries the base current which is $1/(B + 1)$ as large as the emitter current feeding a bit line. Consequently the power demand for a decoder decreases by a few orders of magnitude. MOSTs (Fig. 10.26b) enable a still lower power demand since the gate circuit does not draw current at all. The use of MOSTs offers additional possibilities for efficient application of ROMs.

First, the data can be inserted with the use of a "subtler" approach, rather than resorting to lead burning. The customer receives the homogeneous ROM matrix (with MOSTs in all nodes) as a semifinished product, without gate metallizations. He performs the last photolithographic operation using such a photomask that provides

for metal gates only on those transistors which must transfer a 1 to the bit line. The remaining transistors will have no gates and so will be idle.

Second, the use of MOS transistors allows evolving *semipermanent* or *programmable* read-only memories (PROMs), in which it is possible to refresh the stored information from time to time. Such a possibility is very useful for the customer despite the fact that its realization involves certain manufacturing difficulties. The general principle that underlies the PROM comes to ensuring a *reversible change in the MOST threshold voltage*. Thus, if the condition achieved is such that $|V_0| > V_{AL}$ (see Fig. 10.26b), the address pulses will fail to drive a transistor on; the transistor will stay *inactive as if it were absent*. At the same time, other transistors in which $|V_0| < V_{AL}$ will function normally.

Two methods are presently available for control of the threshold voltage. Both depend on the introduction of additional charges into the dielectric. The first uses MNOS transistors described in Subsec. 7.8.4. The charge build-up and removal are effected by short high-voltage pulses of opposite polarity fed to the gate (see Fig. 7.33b). The second method, applied to memory cells using "conventional" MOS transistors with a **single-layer** dielectric, consists in the following. A sufficiently high voltage impressed on the gate initiates an avalanche breakdown in the dielectric, which thus stores up electrons. The threshold voltage changes accordingly. The electron charge remains unchanged for a rather long time as it does in MNOS transistors. The charge can periodically be restored (regenerated) if need be. To rewrite the information requires *expelling the electrons* from the dielectric. For this, the chip is illuminated with ultraviolet light which causes a photoelectric effect (knockout of electrons from the dielectric).

10.8. Large Scale Integrator

The tendency toward a higher scale of integration was evident from the very first days of microelectronics. Initially, every package was used only for one IC logic element. A later approach was to use a multipin package for the assembly of a few such elements. This reduced the total number of packages in apparatus but did not lead to any new stage of development. The breakthrough came with adoption of the metallization technique for interconnecting simple ICs disposed on a single chip into complex functional blocks. In that period, *medium-scale* integration (MSI) evolved first and then *large-scale* integration (LSI) began to appear. It may be said that *integration of simple ICs lies at the basis of LSI*.

10.8.1. General characteristic of LSI. Logic gates of the RTL, TTL, and other types are a classical example of simple, or small-scale integration (SSI). JK flip-flops consisting of 8 to 10 logic gates occupy an intermediate position between SSI and MSI circuits. MSI refers to small subsystems such as adders, counters, RAMs and ROMs ranging in size from 256 to 1 024 bits; LSI refers to complete logic systems such as memories with 4 kbits and over, arithmetic-logic and control units of computers, and digital filters. The highest scale of integration is inherent in **homogeneous** structures—memories—which show a packing density of around 100 000 elements on a single chip. These ICs represent *very large scale integration* (VLSI).

LSI circuits give a sharp improvement in all basic parameters in comparison with analogous functional systems composed of individual ICs. Indeed, the integration of circuits on a single chip allows a reduction in the number of enclosures, assembly and packaging operations, and external connections (which are the least reliable). These factors naturally aid in decreasing the size, mass, and cost of ICs and in improving reliability. Additional advantages of the larger-scale integration include: a smaller number of bonding pads, and hence greater space saving; reduced length of connections, and thus a higher speed of response and improved noise immunity¹; and lower spread in parameters because all ICs are disposed on one chip and produced in a single manufacturing process.

A higher scale of integration can be achieved in two ways: *by increasing the packing density* (that is, decreasing the area of elements and also the interconnection pattern area) and *by increasing chip sizes*. Both of the approaches call for solving a number of complex technological problems. First of all, it is necessary to increase lithographic resolution, stabilize fabrication procedures, secure adequate hygiene of work, and reduce the density of silicon surface defects. Some problems involving the design and technology control of LSI circuits are discussed in the next subsection.

Note that transition from simple ICs to a large IC does not come merely to implementing the requisite pattern of interconnections, leaving the individual structure of each simple IC intact. A typical feature of modern LSI circuits is the so-called *physical or functional integration*. This feature implies that one and the same structural region of an IC must perform a few functions. I²L circuits may be taken as an illustrative example (see Fig. 10.10). In these circuits, the epi-*n* layer acts as a base for the *pnp* transistor and at the same time functions as an emitter of the *nnp* transistor and

¹ With conversion from SSI to LSI, the average length of interconnections on the printed circuit board decreases from 100 to 1 cm, and an average delay in interconnections (at a signal velocity of 10^{10} cm/s) diminishes from 10 to 0.1 ns.

the base of the *npn* transistor is at same time the collector of the *pnp* transistor. Functional integration enables a substantial increase in the packing density because it obviates the need for many isolating islands and interconnections.

Conversion to LSI has raised the number of microcircuitry problems which are in no way less severe than manufacturing ones. Moreover, both categories of problems are intricately interwoven: *the questions "how to do" and "what is to be done" that arise in evolving LSI designs must be treated as an integral problem.*

One of the first-priority problems facing the LSI development engineer is the problem of securing the technically and economically warranted complexity of an LSI circuit. It is necessary to combine sufficient **complexity** (to achieve the best advantages offered by large-scale integration) with sufficient **versatility** (to ensure the economically justifiable production run). Experience shows that such a compromise is possible to achieve by implementing *elements redundancy and multifunctionality*.

Really, if the number of ICs on the chip is more than necessary to perform a **definite** function, then the same set of ICs can be used to realize any LSI configuration of whatever functional complexity by merely varying interconnection patterns. Redundancy also permits varying the functions of the same LSI circuit by *changing electrically interconnections of the ICs incorporated into it*; this changing is accomplished in accordance with individual programs. This approach is typical of modern LSI circuits; it is referred to as a way that makes it possible *to replace software by system means*, that is, to carry out programming directly with the aid of an LSI circuit. It is exactly this principle that underlies the structure of LSI microprocessors; these devices together with LSI memories form the basis of modern digital engineering.

The ROM shown in Fig. 10.25 gives a good example of multifunctional operation. Assume that for the selection of address lines $AL_0 \dots AL_3$ it is necessary to apply respective binary codes 0000, 0001, 0010, 0011 (that is, numbers 0, 1, 2, 3) to the decoder input. Also, assume that the respective matrix rows contain information as binary codes shown in Fig. 10.25 (that is, numbers 3, 4, 7, 12). So applying, for example, a digit 2 level to the input gives a digit 7 at the output, etc. It is easy to see that the ROM implements the function $y = x^2 + 3$. Writing other sets of numbers along the rows allows performing another function.

10.8.2. Problems of raising the scale of integration. Experience gained in LSI development has revealed that there is a number of problems which impose limitations on the packing density and need to be solved one way or another to speed up the progress in microelectronics.

Heat removal problem. With given sizes of elements, a larger scale of integration can be achieved by increasing the packing density, that is, by bringing the elements on the chip closer together. But a higher packing density inevitably results in a greater power dissipated per unit area. In modern structures of silicon ICs, the permissible specific power for a chip does not exceed 5 W/cm^2 .

This means that the permissible power for a chip 4 mm^2 in area is not over 200 mW . With an average power of 5 mW per IC logic element, the given chip can accommodate not more than 40 logic elements.

A natural way of overcoming this limitation is to use transistors and circuits adapted for work in the **microampere** region. For example, in order for 1 000 gates to be disposed on the same area of 4 mm^2 ,

they should have a dissipation power of not over 0.2 mW . These are I^2L and CMOS gates (see Table 10.1).

It may certainly happen that, whatever the type of logic element employed, a given chip area cannot allow for the desired scale of integration. Then, one has to resort to the chips of a larger area. In principle, this approach offers considerable scope but is not free of limitations either.

The causes of limitations are

dislocations inevitably present on the surface of a semiconductor (see Sec. 2.2). Any dislocation within the active area makes a transistor or an IC unsuitable for use and so the entire LSI circuit may become defective. That is why an increase in chip sizes entails a higher percentage of rejected circuits and thus a lower yield of LSI chips (Fig. 10.27).

Technological advances enable a considerable decrease both in dislocation density and in element sizes. Either of these factors contribute to a higher yield in the given range of chip sizes. However, the chip area is limited at each manufacturing stage by an economically justifiable yield. Thus, if a 5% yield is acceptable, then, as follows from the curves 2 of Fig. 10.27, $S \leq 22 \text{ mm}^2$ for bipolar LSI chips and $S \leq 33 \text{ mm}^2$ for MOS LSI chips. The permissible power dissipation for chips of these sizes lies between 1 and 1.5 W .

An attempt is sometimes made to cool LSI chips or their substrates artificially. But this approach is not versatile enough and far from economical.

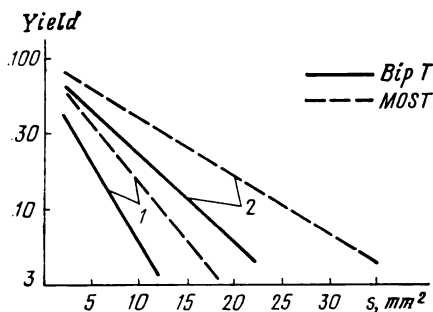


Fig. 10.27. Limitations on the chip area

1—1974; 2—1978

Metallization problem. The internal structure of LSI circuits is so intricate that the designer is unable to work out the element layout for a reasonably accepted time and also develop the optical metallization pattern that would feature a minimum total length of interconnections and, besides, have no crossovers. For this he must compare thousands of versions and prototypes. The task can only be handled with the aid of a computer processing the data according to a specially developed program.

Experience shows that in most LSI circuits it is impossible to produce interconnections in **one** plane without crossovers. For this reason, LSI designs employ *multilevel metallization*, commonly in two or three planes. Isolation of the layers from each other and required connections between the metallizations at various layers present a specific problem involved in the fabrication of LSI circuits.

Both problems mentioned above—programmable computer-aided interconnection layout design and multilevel metallization technology—can presently be regarded as solved problems. One more important problem still awaits its complete solution. It can be formulated as an alternative: whether metallization should be conducted before or after a check of LSI circuits for serviceability. Consequently, in use so far are two methods of carrying out metallization.

Method of fixed metallization. In this method, metallization is conducted before LSI circuit checking with the aid of one photomask providing a definite interconnection pattern. If the circuit has any defects, these can only be detected in the subsequent check, and so a substantial share of all the metallizations may turn out to be made in vain.

Method of selective metallization. This method uses the so-called base chip containing more simple ICs than necessary for implementing the desired functions. For example, if nine ICs are needed to perform a certain function, the base chip will have 12 or 16 ICs. Before metallization, **all** the ICs incorporated into the LSI circuit are checked and then the *control chart for defects* is drawn. A **set** of photomasks is usually available to execute a few variants of metallization depending on which of the ICs are **good**. Choosing the right photomask with the aid of the control chart and “passing” by the faulty ICs during metallization make it possible to fabricate quite an adequate LSI circuit.

The base chip method ensures a substantially higher yield. But it involves additional expenses for the complete control of ICs and for designing of the photomask set. Besides, in order to carry out the check on every IC, its terminals must be made complete with bonding pads which occupy a rather large area.

Problem of performance control. An electrical check on the parameters of an LSI circuit before its packaging is carried out with measuring probes pressed against the bonding pads expected to be

external terminals. Probes are thin metal wires with a point 5 to 10 μm in diameter. A few probes make up a probe head, a kind of wire "brush" in which each probe is in contact with a corresponding bonding pad measuring 100 to 100 μm . The number of external terminals in LSI circuits is much larger than that in simple ICs because the former have to perform functions of a greater complexity. This number can be as large as 32 to 64 and over. As an illustration, assume that a circuit has 50 terminals and consider that two values (0 or 1) are possible at each output. Then, the complete checkout of the LSI circuit for functionality (under static conditions only)

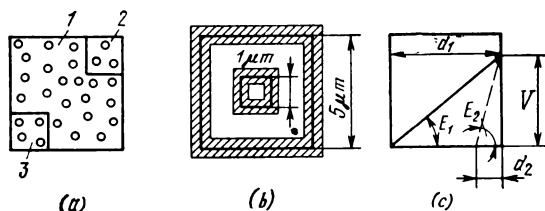


Fig. 10.28. Limitations on the minimum sizes of IC elements

(a) impurity distribution fluctuations; (b) manufacturing tolerances; (c) field strength growth

would require $2^{50} \approx 10^{15}$ measurements. If it takes 1 μs to make one measurement, the entire control procedure for one LSI circuit will take about 25 years.

Consequently, apart from control *automatization*, there is also the need to simplify quality control *procedures*. Measurements should be selective of necessity; the number of measurements made to attest to the serviceability of an LSI circuit (to a definite probability) commonly ranges from 200 to 300.

The choice of the parameters to be checked, the sequence and rules (algorithms) of testing, and also the development of requisite measuring devices and computer-aided quality control programs often present a problem which is no less complex than the design of an LSI circuit itself.

Physical limitations on element sizes. In modern LSI circuits, the sizes of active regions come to 2-5 μm , and there is a tendency toward further scaling down. The process of scaling down in element size, however, is not free from some principal limitations. Fig. 10.28 illustrates them in simplified form.

First, with a decrease in area, a *nonuniform* (statistic) *impurity distribution* in the semiconductor begins to make itself felt. Let the squares in Fig. 10.28a represent the configuration of an emitter layer. With a large area (square 1), the numbers of impurity atoms in two identical squares will practically be the same. With a small area (squares 2 and 3), the numbers of atoms may vary noticeably

(three and four atoms in Fig. 10.28a). Correspondingly, the impurity concentrations in emitters and hence the injection efficiencies will be different too [see Eq. (4.22)]. Analysis shows that this factor becomes substantial with a square side of less than 1 μm .

Second, with a reduction in linear dimensions, the role of *manufacturing tolerances* becomes more important (Fig. 10.28b). Thus, if the photolithography error is $\pm 0.2 \mu\text{m}$, then, taking the linear dimensions to be 5 μm (a large square), the areas of elements will differ insignificantly (by 20%), but with the dimensions being 1 μm (a small square), the areas will differ by a factor of 2.3.

Third, with a decrease in linear dimensions, *the electric field strength* in semiconductor layers rises (Fig. 10.28c). At the same voltage $V = 0.2 \text{ V}$, the field strength in a layer 5 μm thick is comparatively low (400 V/cm), but grows to 10^4 V/cm in a layer 0.2 μm thick and thus *exceeds the critical field strength* (see p. 48). The semiconductor layer then exhibits nonlinear properties.

It may also be shown that at linear dimensions below 1 or 2 μm , such factors as noise fluctuations, cosmic radiations, and the earth's radioactive background begin to play a definite role.

Considering that conventional photolithography offers resolutions only within 0.7-1 μm , the region below 1 μm in size may be said to be a critical one as regards all the physical and technological aspects involved. We thus have ground to consider "submicronic microelectronics" as an independent scientific and engineering field.

10.8.3. Hybrid LSI. This type of LSI is not an alternative of semiconductor LSI. Rather, it may be regarded as an adequate **design** evolved in developing modern microelectronic apparatus.

The basic difference between simple hybrid ICs and hybrid LSI circuits consists in that the former use simple uncased transistor and diode chips as active components, while the latter use uncased IC and LSI chips. Therefore, hybrid LSI circuits can realize more complex functions than semiconductor LSI circuits. Like semiconductor LSI circuits, large hybrids often employ multilayer metalization.

As noted in Sec. 7.11, the tendency today is to abandon the use of film components, primarily capacitors, in the ICs. In large hybrid ICs, this tendency is still more noticeable. Hybrid LSI circuits most often contain **only metallizations** and active components in the form of ICs and LSI circuits. So, the notion of a large HIC (which presupposes the presence of film passive elements) often reduces to the notion of a thin-film or thick-film *commutating board* whose main function is to unite a number of ICs and LSI circuits into a single functional block known as a *microassembly*.

A commutating board is a microelectronic analog of a printed circuit board which up to now has been a basic constructional unit

of radioelectronic devices. As regards a microassembly, its qualitative distinction from the units mounted on printed circuit boards consists in that it represents a complete *electronic device* (a supercomponent of electronic circuits) furnished with its own package and characterized by definite specifications. The functional complexity of such an electronic device is much greater than that of LSI and even VLSI circuits. Microassemblies used as “supercomponents” can be mounted on a printed circuit board to make up supercomplex blocks of an apparatus and often the entire setup.

10.9. Charge-Coupled Devices

The *charge-coupled device* (CCD) is an array of **interacting** MOS structures. Interaction is due to the common semiconductor layer and small spacings between the MOS structures (Fig. 10.29).

A CCD operates on the principle of building up a **local space charge** of minority carriers—*charge packet*—in some MOS elements and transferring it along the surface from one MOS structure to the other by duly varying voltages on metal gates.

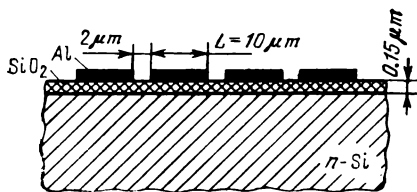


Fig. 10.29. CCD structure

Because the common semiconductor layer is essential to its operation, the CCD may formally be regarded as a specific *semiconductor device* which, like a transistor, cannot be built from (or even modeled with) discrete components. But since it consists of

a great number of technologically combined MOS structures spaced at rather small intervals, the CCD may be thought of as a typical microelectronic device, that is, an *integrated circuit*. Moreover, the CCD is an example of the **large-scale** integrated circuit because it can comprise a few thousand MOS structures.

10.9.1. Basic processes. By analogy with MOS transistors, metal electrodes in the CCD are called *gates*. The operating voltages on CCD gates are higher than the threshold voltage¹. As a result, they produce comparatively deep depletion layers in the semiconductor under the gates. The formation of thin inversion layers near the semiconductor surface (see Fig. 2.22c) is undesirable here. The explanation will be given below.

¹ CCDs most commonly use *n*-silicon, and so the voltages are of negative polarity. By saying “lower” or “higher”, we imply that the voltages are lower or higher in magnitude.

As known, the depth of a depletion layer is directly dependent on the gate voltage [see Eq. (2.53)]. Because the intervals between MOS elements are small, the depletion layers of all elements merge and form a **single** depletion layer whose "bottom" has a definite *relief* corresponding to the distribution of voltages on the gates (Fig. 10.30). Thus, if the voltage $-V_1$ on all the gates is the same, the depletion layer along the entire surface has the same depth (Fig. 10.30a). If the negative voltage on a certain gate is higher than that on the two adjacent gates, a "pit" forms under that gate (Fig. 10.30b and c). The *geometric* relief of depletion layers is in agreement with the *potential* relief: in the region of "pits" of the depletion layer there exists a minimum of potential, known as a *potential well*.

Assume the voltages $-V_1$ on the gates G_1 and G_3 (see Fig. 10.30b) are the same and the voltage $-V_2$ on G_2 is made more negative than $-V_1$. Electric fields then will appear at the boundaries of G_2 , which impede the transfer of positive charges—holes—from under this gate. So, a hole charge packet produced under the gate G_2 in one way or another will be **retained** in this region for a long time. Indeed, the holes cannot leave this region because the retarding fields are present at its boundaries, and the region, now depleted, hardly contains electrons with which the holes could recombine. Then, the CCD is said to operate in the *storage mode* at the storage voltage V_2 .

The **total** positive charge under the gate is determined by the gate voltage (V_2 in the given case). Therefore, *the appearance of a hole packet is attended by a local decrease in the charge of "uncovered" donors in the depletion layer and by a local decrease in the depth of this layer*. In Fig. 10.30b, the relief of the depletion layer *in the absence* of holes is shown by a dash line. It is obvious that the charge of a hole packet reaches its maximum when the depletion layer relief levels off; the retarding fields at the boundaries between the gates

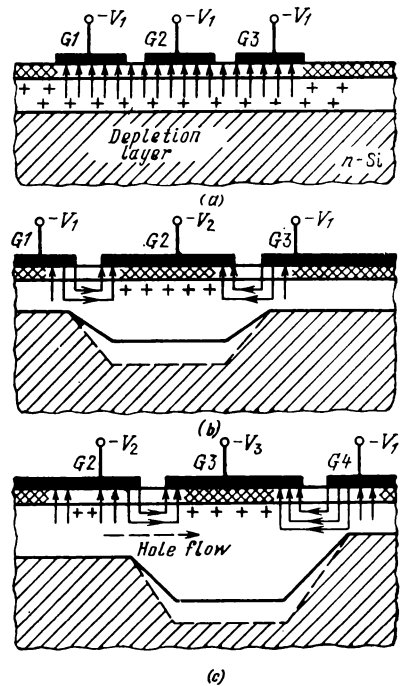


Fig. 10.30. Structure of the depletion layer and electric field in a CCD

(a) quiescent mode; (b) storage mode; (c) transfer mode

then disappear and the hole packet spreads along the entire surface. The maximum permissible charge of a hole packet is given by

$$Q_{\max} = (V_2 - V_1) C_0 (ZL) \quad (10.29)$$

where C_0 is the capacitance per unit area of the insulator [see Eq. (5.1)], Z is the gate width, and L is the gate length (see Fig. 10.29).

We have mentioned earlier that the formation of hole inversion layers under the gates is undesirable. Indeed, during the storage of a hole packet under the gate, **additional** holes appear as a result of thermal generation of carriers. The packet charge then grows and becomes comparable to the hole charge under the adjacent gates where thermal generation also takes place. Finally, the charges under all gates level off and the notion of a charge packet that underlies CCD operation becomes meaningless. Consequently, *the storage time has an upper limit*. This limit depends on the change in the charge of a hole packet allowable during storage. If the permissible variation is 1%, then the storage time does not usually exceed 10 to 20 ms.

Thus, the CCD operation relies on the nonequilibrium conditions of MOS elements, and the device itself is of the *dynamic* type.

Consider now the process of transfer of a charge packet from gate to gate. Set $-V_3$ on G_3 be made more negative than $-V_2$ on G_2 (see Fig. 10.30c). An **accelerating** field will then build up at the boundary between G_2 and G_3 . This field will aid in transferring holes to G_3 . So, the hole packet stored under G_2 will move toward G_3 and remain under the latter because at the boundary of the next gate G_4 a **retarding** field impedes its further motion.

The CCD is said to operate in the *transfer* (write) *mode* when the charge packet moves from under one gate to the other. The voltage V_3 is called the *transfer voltage*.

The full transfer of the charge toward the adjacent gate does not take place because of *charge loss*. This charge loss is due to two causes. First, the process of charge flow from gate to gate is asymptotic in character, and so not all the charge is able to move to the adjacent gate in one transfer step. Second, a fraction of carriers stored under the preceding gate drop into surface traps and have no time to escape them during transfer. In order for the charge losses to be at a minimum, it is *necessary* that the charge transfer (write) time should be sufficiently long. It commonly reaches 50 ns.

It will be readily perceived that, other things being the same, the transfer time decreases with a decrease in the spacing between gates and with an increase in the carrier mobility and transfer voltage.

10.9.2. CCD parameters. To ensure the storage and transfer of charge packets it is necessary to change voltages on the gates in strict sequence.

Figure 10.31 shows a typical *three-phase control circuit* of the CCD and also one of the methods for insertion and extraction of the nonequilibrium hole charge with the aid of *pn junctions*. The voltages of phases *A*, *B*, and *C* are applied in sequence to each third gate of the device (Fig. 10.31a) and shifted in time by 1/3 period (Fig. 10.31b). The voltage V_1 is commonly a **constant bias** applied

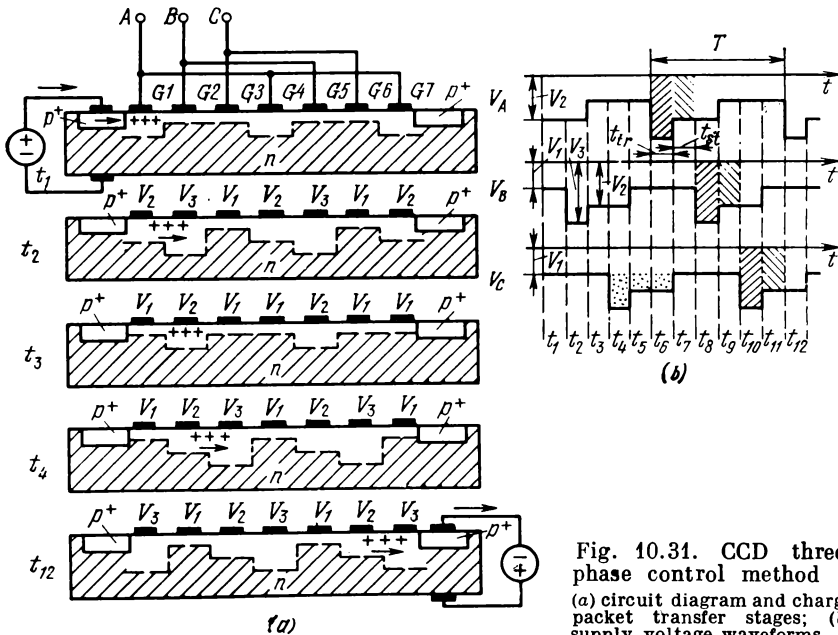


Fig. 10.31. CCD three-phase control method
(a) circuit diagram and charge packet transfer stages; (b) supply voltage waveforms

to all gates, and V_2 and V_3 are voltages resulting from superimposition of additional pulses on this bias (the pulse waveforms are shown by dots on curve V_C in the interval $t_4 \dots t_8$).

Assume that the gate voltages in the interval t_1 are such as shown in Fig. 10.31b. Also, assume that at the start of this interval holes have been injected through the input p^+n junction under the first gate (for this, it is necessary to apply a forward voltage pulse). The injected holes will be kept under the first gate because its voltage is more negative than that of the second.

In the interval t_2 , a transfer voltage V_3 is applied to line *B*. As this takes place, the holes move from the first to the second gate. In the interval t_3 , the voltage on line *B* decreases to V_2

corresponding to the storage mode. Simultaneously, the voltage on line A drops from V_2 to V_1 . This prevents the holes from coming back under the 1st gate. In the interval t_4 , when V_3 is applied to line C , the charge is shifted from under the 2nd to the 3rd gate. The process goes on further in a similar way.

In the interval t_{12} , the voltage V_3 is passed to the 7th (last) gate. Since the output p^+n junction is **reverse** biased, the holes travelling from the 6th to the 7th gate are instantly drawn by the junction field to give a current pulse to the output circuit. The transfer of the charge injected in the interval t_1 is now completed. It is certainly possible to inject new hole packets through the input p^+n junction during transfer of the charge.

The typical values of storage and transfer voltages (V_2 and V_3) range from 10 to 15 V and from 20 to 25 V respectively. The bias voltage V_1 is near the threshold voltage of MOS elements (2 to 4 V).

From Fig. 10.31*b* it is seen that the period T of each phase is the sum of three transfer intervals t_{tr} and three storage intervals t_{st} (all shown hatched in the figure for clarity). Thus,

$$T = 3(t_{tr} + t_{st}) \quad (10.30)$$

Knowing the number N of MOS elements, it is not difficult to determine the overall delay t_d as the pulse goes from input to output. Since the delay between the two **adjacent** elements is $1/3 T$, then, multiplying this value by $N - 1$, we obtain

$$t_d = 1/3 (N - 1) T = 1/3 (N - 1)/f \quad (10.31)$$

where $f = 1/T$.

In practice, t_{tr} and t_{st} are generally not equal. The relation between them is different depending on the purpose a CCD has to serve.

The frequency is a maximum if $t_{st} \ll t_{tr}$:

$$f_{\max} = 1/(3t_{tr}) \quad (10.32)$$

The transfer time must be long enough to transport the charge from element to element as fully as possible. The causes of incomplete transfer were explained above.

The qualitative characteristic of charge transfer is the *transfer efficiency*

$$\eta = 1 - \Delta Q/Q = 1 - \varepsilon \quad (10.33)$$

where Q is the charge packet transferred, ΔQ is the net charge lost from the packet in a transfer cycle, and ε is the *loss factor* (transfer inefficiency). If η is equal to $1 - \varepsilon_1$ for a spacing between two adjacent elements, then for the array of N elements constituting the CCD the transfer efficiency is near $1 - N\varepsilon_1$.

So the permissible number of elements depends on the loss factor ε_1 . The latter in turn heavily depends on the spacing between

elements and on the duration of a transfer pulse. At a spacing of 2 or 3 μm and t_{tr} of 20 to 50 ns, the loss factor ϵ_1 is 2×10^{-4} to 5×10^{-4} , which permits using a few hundred elements.

One of the ways of reducing the loss factor is to neutralize the effect of traps (see above), in particular, by injecting a *background charge* into the CCD to fill the traps and prevent the holes of the **working** charge packet from falling into them. The use of a background charge decreases the loss factor by about one order of magnitude.

A maximum frequency $f_{\max} = 6$ to 15 MHz corresponds to $t_{tr} = 20$ to 50 ns. At a maximum working frequency (when $t_{st} \ll t_{tr}$), *the storage period is in essence nonexistent: the charge is uninterruptedly transferred from one MOS element to the other.*

A minimum frequency corresponds to the reverse condition $t_{st} \gg t_{tr}$:

$$f_{\min} = 1/(3t_{st}) \quad (10.34)$$

We have mentioned earlier that the storage time has an upper limit: it must be so small that during transfer of charge Q through the **entire** CCD for the time $(N - 1)t_{st}$ the stored parasitic charge cannot exceed fractions of Q . Thus, if $Q_{par} = 0.2Q$ and $N = 200$, then the parasitic charge during its storage in the potential well must not be in excess of 0.1% of the useful charge. Commonly, $t_{st} \leq 1$ to 10 ms and, correspondingly, $f_{\min} \geq 30$ to 300 Hz.

A CCD shows an advantage in that it draws little power. Indeed, the CCD does not practically consume power in the storage mode. Currents flow through the gates only when the charge transfer takes place. So the maximum power consumed during transfer of one charge packet, taking into account Eqs. (10.29) and (10.32), will take the form

$$P_{\max} = \frac{Q_{\max}(V_3 - V_2)}{3t_{tr}} \approx (V_3 - V_2)^2 ZLC_0 f_{\max} \quad (10.35)$$

For typical values of $V_3 - V_2 = 10$ V, $Z = 20$ μm , $L = 10$ μm , $C_0 = 200$ pF/mm², and $f_{\max} = 10$ MHz, we get $P_{\max} = 4$ $\mu\text{W/bit}$.

10.9.3. Applications and types. The delay of an input pulse for an exactly specified time t_3 is one of the important functions of a CCD.

A second function is related to a possibility of a comparatively long storage of information. For this it is enough to interrupt the sequence of control (clock) pulses after the packets of injected holes have filled the corresponding MOS wells. During readout, clock pulses are again fed to the CCD and the written information is transferred **in steps** to the output. Unlike the address matrix system of Fig. 10.23, the CCD of this type does not allow for **random** access. All the same, it holds an appropriate place in digital engineering;

the capacity of such a CCD is rather large, 8 to 16 kbits and more. To ensure **prolonged** storage, periodical *regeneration* of written information is necessary. The procedure is the same as for other dynamic memories (see Fig. 10.24b).

A unique feature of the CCD is that the hole charge can be introduced not only with the aid of a *pn* junction but also by **locally illuminating** the surface. As a result, a charge proportional to the illumination intensity builds up under the corresponding gate. The light causes generation of electron-hole pairs. The gate field repels

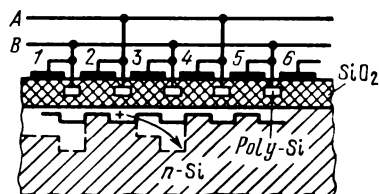


Fig. 10.32. Two-phase CCD

electrons, thereby enabling accumulation of holes in the potential well. If the intensity of illumination is different in various portions, then a combination of charges under the gates will characterize an image projected on the CCD. Applying a control three-phase voltage to the CCD gives a train of pulses at the output, the amplitudes of which are proportional to the illumination intensity in various regions (the principle that finds wide use in television). The three-phase control method (Fig. 10.31) is a historically first and, to a certain extent, classical one. But it suffers from the following disadvantages:

1. Three adjacent MOS elements (Fig. 10.31a) are needed to store one hole packet, which makes impossible a more efficient use of the area.

2. The metal plates of each phase must be arranged in their own plane to exclude crossovers, and so three-layer metallization is necessary.

3. Close mutual arrangement of the elements (2 or 3 μm) is fraught with "shortings", that is, rejection of the device. A further decrease of spacings (to raise the scale of integration) is so far impracticable.

The structure of a CCD shown in Fig. 10.32 is more perfect. First, it is a two-phase CCD, so the device needs only two MOS elements to solve and transfer one hole packet, thereby permitting the use of two layers of metallization. Second, it has no spacings between elements. In this structure, each MOS element contains two interconnected gates, one being a "buried" silicon gate (see Fig. 7.31b) and the other a conventional aluminum gate located on the oxide layer surface. Since silicon gates lie closer to the semiconductor than aluminum ones, *the depletion layer depth proves different within*

the confines of one element. This difference in depth remains in the transfer mode too (see the dash line in Fig. 10.32). That is why the transferred charge cannot come back despite the two-phase supply. This structure is attractive in that it allows for a greater density of elements and higher degree of their integration, and also features an increased speed ($f_{\max} = 20$ to 50 MHz).

The structure of a CCD with a buried channel shown in Fig. 10.33a ensures a still higher speed. The n substrate has an epi- p layer grown to a thickness of a few micrometers. The potential distribution,

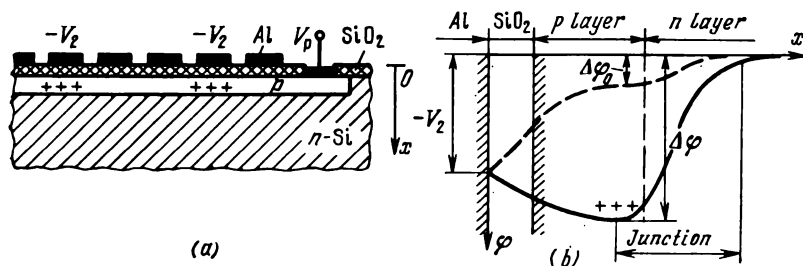


Fig. 10.33. Buried-channel CCD

with the gates reverse biased and at $V_p = 0$, is shown in Fig. 10.33b by dash lines. A sufficiently high negative voltage $-V_p$ applied to the p layer will maintain a reverse bias on the pn junction, close in value to V_p , and so the potential distribution will be such as shown in Fig. 10.33b by solid lines. As seen, the potential minimum has shifted from the boundary of the dielectric into the p layer bulk. It is exactly the region which the holes will now occupy.

So in the buried-channel CCD, the charge packets are isolated from the surface and located in the semiconductor bulk. Hence, the carrier mobility is increased and the effect of traps near the surface eliminated. Both of these factors aid in increasing the speed and decreasing the loss factor. The maximum working frequency for the buried-channel CCD reaches 500 to 800 MHz and the loss factor is 10^{-6} to 10^{-7} . This type of CCD may include a few thousand MOS elements. But since the potential well in this structure lies away from the surface, the device requires higher operating voltages and shows a lower maximum charge in the packet than the CCD with a surface channel.

10.10. Operational Amplifiers

In comparison to digital devices, analog counterparts feature a greater diversity as regards the kind of signals handled, functions performed, and also the purposes served and types of internal structure.

Therefore, unification of the building blocks in the field of analog devices is possible only proceeding from *multifunctional* units.

At present, it is customary to regard amplification, comparison, limiting, multiplication, and signal frequency filtering as *main analog functions*. Each of these functions is generally performed by analog ICs of a particular class. But all these specialized ICs principally come from the basic, most versatile and multifunctional unit called an *operational amplifier* (op amp) with which we shall deal in this section.

10.10.1. General characteristic. The operational amplifier is generally a dc amplifier with a differential input and single ended output, the most important characteristics of which are a very high gain, a high input and a low input resistance. The explanation of the notions "high" and "low" will follow later in the text. The standard op amp schematic symbol is illustrated in Fig. 10.34.

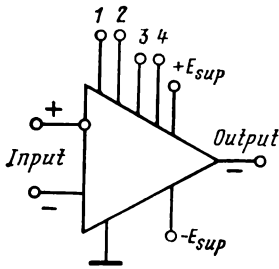


Fig. 10.34. Standard op amp schematic symbol

A signal must not necessarily be differential; it may be applied to one of the op amp inputs with the second input grounded. One of the inputs is called *inverting* and the other *noninverting*, depending on the polarity of signals at the output and input (see Fig. 10.34). As in logic gates (see Fig. 10.1), the inverting input is represented by a circle. In practice, it is most

common to apply a large amount of feedback to an op amp. It is in combination with feedback paths that the op amp is capable of executing a variety of *operations*, hence the name operational amplifier. A typical op amp circuit with resistive negative feedback appears in Fig. 10.35, where the resistance R_1 includes a signal source resistance R_g . In this circuit version, the op amp performs the function of stable amplification.

Let us idealize the op amp by setting $R_{in} = \infty$ and $R_{out} = 0$ (we shall explain below why such idealization is permissible). It may then be assumed that $I_{in} = 0$ and $V_{out} = -K_0 V_{in}$. The currents I_1 and I_2 prove equal. Write the equality $I_1 = I_2$ in the form

$$(1/R_1) (E_g - V_{in}) = (1/R_2) (V_{in} - V_{out})$$

Substituting $V_{in} = -(1/K_0) V_{out}$, dividing both sides by E_g , and making simple transformations, we obtain the *circuit gain*

$$K = \frac{V_{out}}{E_g} = - \frac{R_2/R_1}{1 + (1 + R_2/R_1)/K_0} \quad (10.36)$$

If the gain of an op amp is sufficiently high, the second term in the denominator of (10.35) may be neglected. Then,

$$K = -R_2/R_1 \quad (10.37)$$

Expression (10.37) is a fundamental one for an op amp. It shows that *under definite conditions the circuit gain depends only on the parameters of a feedback circuit and does not depend on the parameters of the op amp proper*. In particular, the circuit gain is independent of temperature, supply voltage, and changes of factors β , whatever the causes of these changes can be. Replacing resistances R_2 and R_1

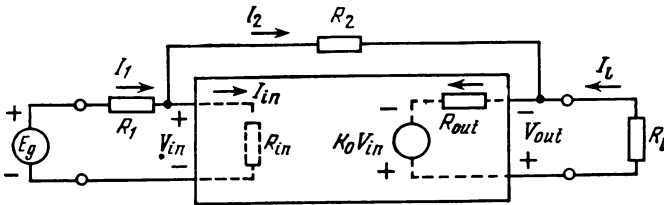


Fig. 10.35. Typical circuit diagram of an op amp

by complex impedances, we can obtain the desired transient and frequency characteristics, which are also independent of the op amp parameters (see p. 442).

Let us specify the conditions under which Eq. (10.37) is true. First of all, it is high values of K_0 that make this expression valid; namely, according to Eq. (10.36), the inequality

$$K_0 \gg (R_2/R_1) + 1 = K + 1 \quad (10.38)$$

must be met. Consequently, *the gain of an op amp must by far exceed the desired gain of the circuit*. For example, if it is desirable that K will be 100, K_0 must be higher than 10^3 – 10^4 .

With an increase in frequency, the value of K_0 inevitably diminishes, which leads to a disturbance of the inequality (10.38). Therefore, *an increased cutoff frequency of K_0 ensures a wider frequency band within which Eq. (10.37) holds good* and offers all its advantages.

At the beginning of the analysis we have disregarded output and input resistances. If we now allow for the finite value of R_{in} , the currents I_1 and I_2 will differ by a value of $I_{in} = V_{in}/R_{in}$. It is easy to show that this correction leads to an additional term R_2/R_{in} in the parentheses of Eq. (10.36). This summand and hence the effect of input resistance will be insignificant if

$$R_{in} \gg R_1 \quad (10.39)$$

The inequality (10.39) limits the permissible values of R_1 and thus those of R_2 at a given value of R_{in} . Higher values of R_{in} permit the use of higher-value resistors in feedback circuits.

As for the output resistance, its effect comes only to a certain decrease in V_{out} in comparison with the emf of an equivalent generator $K_0 V_{in}$. So, we should replace K_0 used in the above formulas by a somewhat smaller value, K'_0 , which is dependent on the relation between $R_l \parallel R_2$ and R_{out} . Commonly, the following inequality is met:

$$R_{out} \ll R_l \parallel R_2 \quad (10.40)$$

The correction for the value of K_0 does not then exceed 10% and thus may be disregarded. The lower the value of R_{out} , the higher the load capacity of an op amp and the smaller the values of resistors that can be inserted into the feedback circuit.

The integrated op amp shown in Fig. 10.34 has some other terminals apart from the input and output terminals. Two of them are intended for voltages of a dual-polarity power supply, one for grounding, and the rest (1 . . . 4) for connecting auxiliary external circuits.

10.10.2. Basic parameters of op amps. For an op amp to have a differential input, its first stage must be a differential amplifier DA (see Sec. 9.6). Depending on its gain, the first DA can be followed either by a second DA or just by a level shifter and other intermediate stages, which, in the final analysis, have to couple the DA with a high-power output stage. The last stage is practically always connected in a class B push-pull circuit (see Sec. 9.9).

The input DA predetermines the list of op amp parameters. This list is practically the same as for an individual DA. The parameters, considered in detail in Sec. 9.6, include voltage gain K_0 , common-mode rejection ratio K_R , input-offset voltage V_{off} and its temperature drift ϵ_{off} , average input current $I_{in\ av}$, and input-offset current ΔI_{in} . As a matter of fact, in the list are also such parameters as supply voltage E_{sup} , supply current I_{sup} , power P_{sup} , maximum allowable input voltages, maximum allowable output current, and some others.

Input and output resistances are rarely included into the list of basic parameters, but then they can be judged of by the values of input and output currents.

It is usual to characterize the speed of response of an op amp by a parameter $v_{V_{out}}$ known as a *slew rate*; this parameter is measured for a step input of a maximum possible amplitude. In a rarer use are the maximum frequency or *unit-gain* frequency f_1 at which the gain drops to $K_0 = 1$.

The development of integrated op amps took place in definite stages, each of which was noted for original circuitry designs and specific types of semiconductor base. We can now distinguish three stages and three generations of integrated op amps. The averaged parameters of these generations are listed in Table 10.3

All the three generations can be placed into the category of universal op amps. The fourth generation which dates from about 1974 refers to specialized op amps. This means that the parameters listed for the fourth generation in Table 10.3 are not simultaneously met

Table 10.3

Typical Parameter for Integrated Op Amps

Generation	K_0 , V/mV	K_R , dB	V_{off} , mV	ϵ_{off} , $\mu V^\circ C^{-1}$	$I_{in\ av}$, nA	ΔI_{in} , nA	$v_{V_{out}}$, V/ μs	P_{sup} , mW	E_{sup} , V
1st	25-50	70-85	2-5	3-5	500	200	0.5-1	10-50	$\pm 5 \dots \pm 20$
2nd	100	80-90	1-2	3	100	20-50	0.5	5-20	$\pm 3 \dots \pm 20$
3d	10^2 - 10^3	90-100	0.5-1	2-3	2-5	0.1-0.2	0.2-2	0.2-50	$\pm 2 \dots \pm 15$
av	10^3	120	0.25-0.5	0.5	0.1	0.005	50-70	0.08-0.15	± 2 to 5
4th									
max									
or									
min	10^5	140	0.02	0.2	0.02	0.001	350	0.025	± 1.5

with in one and the same op amp; they hold true for *various* op amps and are "record" parameters. For example, one type of op amp can have a minimum input-offset voltage but a low speed, the other a maximum speed but high power consumption, and so on.

From Table 10.3 it is clear that the main trends in universal op amps were for higher gain and, primarily, for lower input currents. Decreased input currents aided not only in raising the input resistance but also in improving offset and drift compensation (see Subsec. 9.6.6).

The following parameters of modern universal of op amp can be considered typical:

$$\begin{aligned}
 K_0 &= 10^5\text{-}10^6 & K_R &= 80\text{-}100 \text{ dB} \\
 V_{off} &= 1\text{-}3 \text{ mV} & \epsilon_{off} &= 2\text{-}3 \mu V^\circ C^{-1} \\
 I_{in\ av} &= 5\text{-}50 \text{ nA} & \Delta I_{in} &= 1\text{-}10 \text{ nA}
 \end{aligned}$$

There are various means to achieve substantial improvement in individual parameters of the fourth specialized generation of op amps. Some of these means are described below.

As known, the number of transistors for integrated op amps is not very critical, and so the designer can allow himself to add to an op amp additional elements and stages for improving the amplifier performance. Besides, close location of elements on one sub-

strate favors higher symmetry of DA branches and, hence, better offset and drift compensation. That is why *the parameters of integrated op amps are always substantially higher than for similar discrete circuits.*

10.10.3. Design versions. Every generation of integrated op amps is noted for its own design versions. An example of the simple op

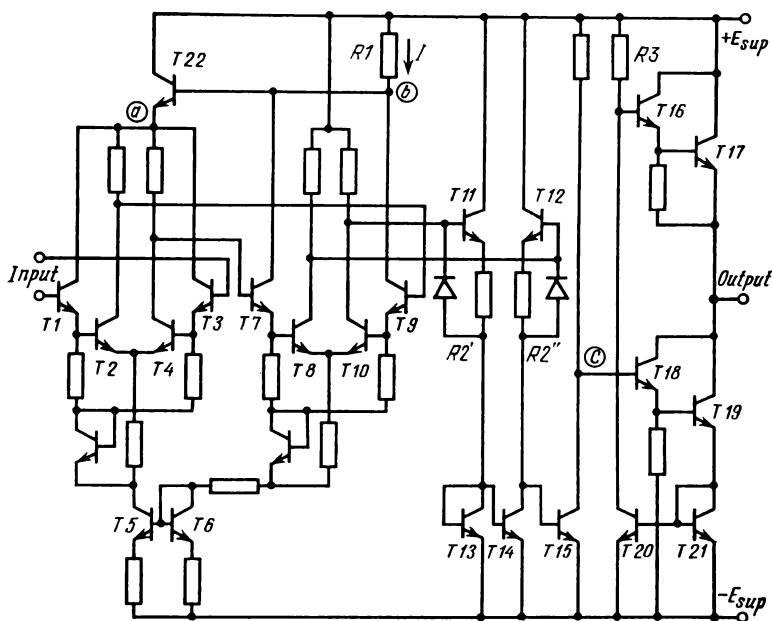


Fig. 10.36. Schematic of the type 1YT402 op amp

amp belonging to the first generation is given in Fig. 10.36. This circuit version uses transistors of only one type (*nnp*) and **resistive** loads in DAs, namely, diffused resistors.

In the schematic diagram we can readily single out the first and the second DA (*T1-T4* and *T7-T10* connected pairwise in the Darlington circuit), level shifting circuits (*T11* and *T12*), intermediate single-ended amplifier *T15*, and a push-pull output stage using Darlington pairs (*T16, T17* and *T18, T19*). The given op amp circuit employs three current reflectors (*T6-T5, T13-T14*, and *T21-T20*). The principle of current reflectors (current mirrors) is described in Sec. 9.11. The first reflector ensures a constant **relationship** between the operating currents in both DAs, the second provides the equality of currents in level shifters, and the third ensures antiphase voltages

at the inputs of the push-pull output stage. A few diodes shown in the circuit play an auxiliary role which will not be considered here.

The transistor T_{22} is connected in an emitter follower configuration. Its purpose is to lower the potential at point a over the supply voltage $+E_{sup}$. Indeed, the potential at point b is equal to $E_{sup} - IR_1$. Since the current I fed to the second DA does not vary, V_b may be considered constant; then $V_a = V_b - V^*$ is also constant and may be regarded as a *supply voltage* for the second DA. Such an artificial decrease in supply voltage makes it possible to use lower-value collector resistors in the first DA, that is, to reduce the time constant of collector circuits and the area for resistors.

The current reflector T_{13} - T_{14} in the level shifting stage enables doubling the signal delivered from the second DA to the input of stage T_{15} . Assume the base of T_{11} has received a signal $+\Delta V$ and the base of T_{12} a signal $-\Delta V$. The increment $-\Delta V$ almost fully goes to the base of T_{15} since T_{12} is connected in an emitter follower configuration. The increment $+\Delta V$ causes an emitter current increment $\Delta V/R_2'$ in T_{11} (and hence in T_{13}).

This increment, reflected by the current reflector to the emitter circuit of T_{12} , produces a voltage drop across R_2'' , which also proves equal to $-\Delta V$ if R_1' and R_2'' are equal in value. So the total change in base potential V_{b15} becomes equal to $-2\Delta V$.

The current reflector T_{21} - T_{20} acts in the following manner. A signal ΔV at point c causes an emitter current increment in transistors T_{19} and T_{21} . This increment repeats in the collector circuit of T_{20} and gives an increment $-\Delta I_e R_3$ on the base of T_{16} . The resistance R_3 is chosen such as to secure an equality $\Delta I_e R_3 = \Delta V$. The voltages on the bases of T_{16} and T_{17} will then be equal but opposite in sign, and this is the condition required for control of the push-pull stage.

Op amps of the second generation feature a number of improvements. First, they use *pnp* transistors along with *nnp* transistors; among other things, this facilitates designing of output stages (see Fig. 9.25). Second, along with simple diffused resistors, the op amps employ pinch resistors showing higher resistance ratings (see Subsec. 7.9.1). Third, along with bipolar transistors, input DAs sometimes use FETs. These transistors are inferior to bipolar counterparts in amplifying and frequency characteristics, but enable a sharp

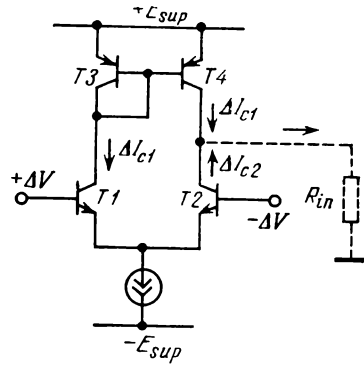


Fig. 10.37. Differential amplifier stage with dynamic load

reduction in input currents and an increase in input resistance. But the main feature of the second generation consists in the replacement of resistive loads in DAs by dynamic loads. The examples of dynamic loads, as applied to MOS transistors, were given above (see Fig. 9.13). As regards bipolar transistors, a typical approach is the use of a current reflector in the collector circuits of a DA (Fig. 10.37).

Since the transistors $T2$ and $T4$ are connected together in the "collector to collector" manner, it is safe to say that the load for $T2$ is r_{c4} and that for $T4$ is r_{c2} . Both these resistances are very high, particularly in the microampere region [see Eq. (4.42)]. The load for both transistors is practically a smaller resistance R_{in} , which is the input resistance of the next stage (shown by a dash line in Fig. 10.37).

Assume the bases of $T1$ and $T2$ have received signals $+\Delta V$ and $-\Delta V$ respectively. The collector current I_{c1} will then change by $\Delta I_{c1} = \alpha (\Delta V/r_e)$. The increment ΔI_{c1} will be reflected to the collector circuit of $T4$ to give a collector potential increment $\Delta V_{c2} \approx \Delta I_{c1} R_{in}$. The increment ΔI_{c2} determined by a signal $-\Delta V$ will give exactly the same value of potential increment. As a result, the value at the output is $\Delta V_{c2} = 2\Delta V (R_{in}/r_e)$, where R_{in}/r_e is the gain. So *the use of a current reflector allows us not only to obtain a high gain (to a few thousands) but also to double the signal at the single-ended output of a DA* (such doubling was illustrated above for a level shifting circuit, see p. 433).

The third generation of integrated op amps typically use super-beta transistors in input DAs (see Subsec. 7.4.4). High values of β , characteristic of these transistors, in conjunction with the micro-ampere operation mode practically offer the same small input currents as the ones provided by FETs, but ensure higher speed and amplification.

10.10.4. Methods of Drift Compensation. Symmetry of an integrated DA helps solve the problem of drift to a considerable degree. But in a number of special applications (particularly in precision measuring devices) this problem remains acute. Consider two methods of improving the temperature stability of devices in integrated form.

The first method relies on stabilizing the temperature of an op amp chip, so that the temperature drift does not exist or, in any case, decreases substantially. The block diagram for temperature stabilization appears in Fig. 10.38.

Apart from the op amp itself, the chip also carries temperature-sensitive elements D (commonly, forward-biased pn junctions), auxiliary amplifier A and heat-liberating elements HE (power tran-

sistors). Write the following relation

$$\Delta T_{IC} = \Delta T_0 + \Delta P_{st} R_t \quad (10.41)$$

where ΔT_{IC} is a change in the chip temperature, ΔT_0 is a change in the environment temperature, ΔP_{st} is an increment in the power of heat-liberating element, and R_t is the chip thermal resistance (see p. 97). Considering that the heat-liberating elements can only heat up the chip, but not to cool it, one can readily conclude that *the given IC always operates at a maximum possible temperature*. This is one of the drawbacks of the method. A second drawback is the need for an additional area (up to 20%) for the temperature stabilization elements.

Denote the temperature sensitivity of D as ε ($V^\circ C^{-1}$), the gain of the auxiliary amplifier A as K_A , and the transconductance for heat-liberating elements as S_{HE} . Then, because of the increment ΔT_{IC} , the power increment ΔP_{st} will take the form

$$\Delta P_{st} = (\varepsilon \Delta T_{IC}) K_A S_{HE} E$$

where E is the supply voltage. Substituting ΔP_{st} into Eq. (10.41) yields

$$\Delta T_{IC} = \Delta T_0 / (K_{st} + 1) \quad (10.42)$$

where the stabilization factor

$$K_{st} = -\varepsilon K_A S_{HE} E R_t$$

It is obvious that the sign of K_A must be opposite to the sign of ε . For typical values of $\varepsilon = -2.2$ mV $^\circ C^{-1}$, $S_{HE} = 40$ mA/V, $E = 10$ V, $K_A = -200$, and $R_t = 0.3^\circ C/mW$, we get $K_{st} = 50$. This means that over the temperature range $\Delta T_0 = 150^\circ C$, the chip temperature will vary by merely $3^\circ C$ (the chip operating temperature will be about $+120^\circ C$).

A radical method of eliminating the effect of drift (but not the drift itself) is to use modulation-demodulation (MDM) amplification instead of *conventional direct amplification*. In conventional amplifiers, slowly varying signals are inseparable from the drift because their frequency spectra coincide. The principle of MDM amplifiers lies in **modulating** a useful signal, that is, shifting its spectrum into a comparatively high-frequency region, while maintaining its amplitude (Fig. 10.39). It is then safe to pass the signal through a conventional amplifier A , being sure that the signal and the drift will not "mix up". At the output of the amplifier the signal is **demodulated**, that is, its initial spectrum is restored.

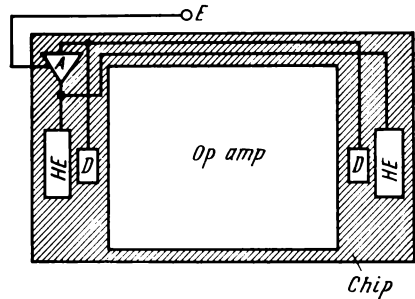


Fig. 10.38. Block diagram for chip temperature stabilization

MDM amplifiers did not gain recognition in discrete circuits because, first, they require additional active elements and also high-quality modulators—choppers. In integrated form, MDM amplifiers prove rather simple; they use MOS switches (see Subsec. 8.7.4), noted for the absence of the residual voltage in the on state, which

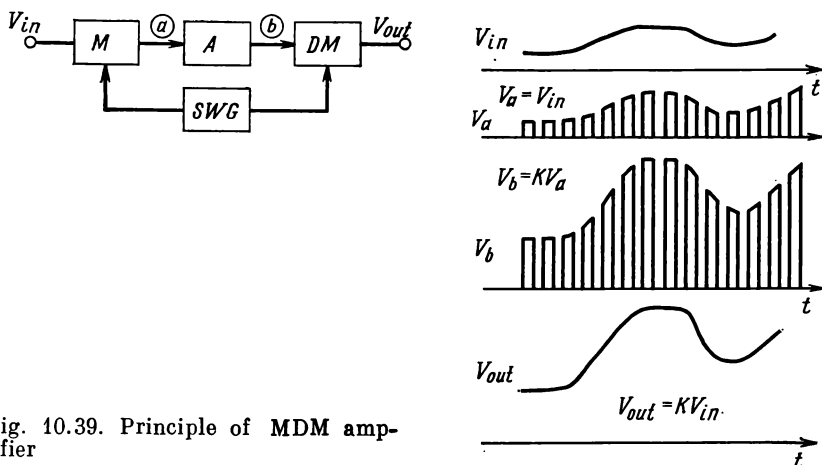


Fig. 10.39. Principle of MDM amplifier

provide high-quality modulation and demodulation. Control over the switches is effected by a square-wave generator disposed on the same chip (SWG in Fig. 10.39).

MDM amplifiers ensure not only a small temperature drift ($0.2 \mu\text{V } ^\circ\text{C}^{-1}$, see Table 10.3) but also a decreased level of 1f noise.

10.10.5. Op amp applications. Prior to considering the examples of op amp uses, let us note that both the analysis and calculation of many circuits containing op amps become simpler under the assumption that *the input voltage of an op amp is equal to zero*. This approach is valid if we know that the op amp in question operates in the *normal linear mode*. In this case the input voltage is a factor of K_0 smaller than the output voltage. The latter does not usually exceed 5 to 10 V. Thus, the values of V_{in} are as a rule equal to merely tens of microvolts and even less, that is, a few orders of magnitude lower than other voltages in the circuits.

In a number of cases, an op amp operates in the **nonlinear** mode. This happens when the gain K_0 ceases to be a proportionality factor for the input and output voltages. Indeed, the output voltage cannot exceed the supply voltage. If we now apply a sufficiently large signal $V_{in} > E_{sup}/K_0$ to the input, the output transistor will either go off or on to saturation (see Subsec. 9.6.5). The output

voltage then takes one of the limiting values, $+E_1$ or $-E_2$, and does not further depend on the input signal.

Fig. 10.40 shows a typical circuit of the voltage regulator. Setting $V_{in} = 0$ (see above), we may write $V_a = V_{ref}$, where V_{ref} is the reference voltage. Expressing V_a in terms of the output voltage of the form $V_a = \gamma V_2$, yields

$$V_2 = V_{ref}/\gamma$$

where $\gamma = R_2/(R_1 + R_2)$. Varying the resistances R_1 and R_2 allows regulating the output voltage.

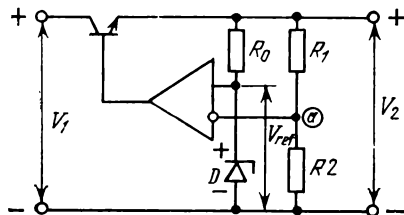


Fig. 10.40. Voltage regulator using an op amp

The use of an op amp makes it possible to solve one of the main problems, namely, to decrease sharply the output resistance of regulator as against that typical of a simple circuit (see Fig. 9.32). To support the point, let us set an increment ΔV_0 at the output.

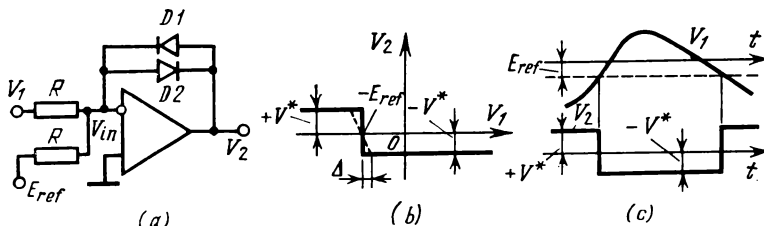


Fig. 10.41. Voltage comparator

(a) circuit; (b) transfer characteristic; (c) comparison function

As it arrives at the transistor base, the amplified increment $K_0\gamma\Delta V_0$ causes an emitter current increment $\Delta I_e \approx K_0\gamma\Delta V_0/r_e$. Dividing ΔV_0 by ΔI_e gives

$$R_{out} \approx r_e/K_0\gamma$$

The **calculated** value of R_{out} can be as low as thousandths of an ohm and less. In this case, the **actual** value is often determined by metallization or conductor resistances. The stabilization factor here proves higher than that of a simple circuit.

Figure 10.41 shows the circuit of a *comparator* whose function is to compare two voltages.

Assume first that $V_1 = -E_{ref}$, where E_{ref} is the fixed reference voltage. With the resistances R being the same, the potential at the inverting input will be a half-sum of V_1 and E_{ref} , that is, equal to zero. Correspondingly, $V_2 = 0$ and both diodes are off. If we now raise the input voltage by ΔV_1 , the potential at the inverting input will become positive and a negative voltage V_2 will appear at the output. The diode $D2$ will then turn on. As known, the voltage on a forward-biased diode is practically a constant value, equal to V^* . Setting $V_{in} = 0$ (see the text at the beginning of the subsection), we arrive at the conclusion: after biasing the diode $D2$ to the on condition, the output voltage is equal to $-V^*$ irrespective of the value of V_1 . If $\Delta V_1 < 0$, the diode $D1$ goes on and the output voltage becomes equal to $+V^*$ and is also independent of V_1 .

The voltage V_{in} at which one diode or another turns on characterizes the comparator *sensitivity* Δ .

This is the ratio of the output voltage V^* to gain K_0 :

$$\Delta = V^*/K_0$$

For example, if $K_0 = 10^5$, then $\Delta \approx 7 \mu\text{V}$. So the output voltage is set at levels $\pm V^*$ when the voltage V_1 deviates from $-E_{ref}$ by an extremely small value. It is impossible to represent the interval Δ on the graph where the scale of V_1 comes to a few volts or to tenths of a volt. Hence, the comparator will produce a **step-like** waveform (Fig. 10.41b). This feature underlies the application of a comparator: it **compares** changing voltages with the reference voltage E_{ref} and senses their equality. Each time the comparator detects the equality of voltages, the output voltage swings sharply to the opposite polarity (Fig. 10.41c). In a particular case where $E_{ref} = 0$, the comparator is referred to as a *zero detector*.

If individual diodes $D1$ and $D2$ are replaced by series diode networks, the comparator output voltage will be higher respectively. But it cannot exceed the limiting values $+E_1$ and $-E_2$ which were mentioned earlier (see p. 436). If we do not connect any diodes at all in the feedback circuit, the output levels will be at $+E_1$ and $-E_2$ and the *sensitivity* will be $+E_1/K_0$ and $-E_2/K_0$. In this case, the voltages V_1 differing from $-E_{ref}$ by a value higher than the response level will correspond to the **nonlinear** mode of operation of the op amp.

Figure 10.42a illustrates a *threshold device* which operates in much the same way as the Schmitt trigger (see Sec. 8.10).

Denote the feedback factor as

$$\gamma = V_0/V_2 = R_1/(R_1 + R_2)$$

Suppose that in the initial state all the potentials (V_1 , V_0 , and V_2) in the circuit are equal to zero. Using the same method as that described in p. 308, we can ascertain that this state is **unstable**: the

smallest fluctuation causes an avalanche-like process. As a result, the output voltage assumes one of the two limiting values, $+E_1$ or $-E_2$ (see p. 436).

Let at $V_1 = 0$ the output voltage be equal to $+E_1$. The potential V_0 will then be equal to $V_0^+ = \gamma E_1 > 0$, and the input voltage (referred to the *noninverting input*) will be given by

$$V_{in} = V_1 - V_0^+ = -\gamma E_1$$

With the control voltage V_1 made negative, V_{in} becomes still more negative, so the output voltage does not change and remains equal to E_1 . At positive values of V_1 lower than V_0^+ , the output voltage will not yet change because the difference $V_1 - V_0^+$ remains negative.

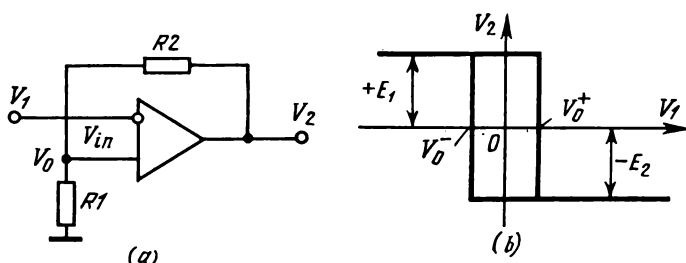


Fig. 10.42. Threshold device
(a) circuit; (b) transfer characteristic

Only at $V_1 \approx V_0^+$ does the input voltage approach zero and the output voltage begin to fall off. The potential $V_0 = \gamma V_{out}$ will then start decreasing along with the input voltage. The avalanche-like process thus sets in, with the result that the output voltage assumes the second stable value, $-E_2$. The potential V_0 now becomes equal to $V_0^- = -\gamma E_2 < 0$. Since the control voltage V_1 has not changed during the avalanche-like process and remained equal to V_0^+ , the input voltage takes the form

$$V_{in} = V_1 - V_0^- = \gamma (E_1 + E_2)$$

A further rise in V_1 has no effect on the value of V_2 .

If now the voltage V_1 decreases and goes through zero to the negative values, the output voltage does not change until after the potential V_1 becomes near V_0^- . Following this, the avalanche process will occur again and the output voltage will swing to E_1 . The function $V_2(V_1)$ is illustrated in Fig. 10.42b. In distinction to the Schmitt trigger, this circuit has a **dual-polarity** transfer characteristic, which is also symmetric about the V_2 axis if $E_1 = E_2$.

The threshold device forms the basis for a wide class of pulse

circuits: square-wave generators, sawtooth generators, pulse shapers, and others.

In conclusion, consider op amp circuit versions with feedback paths (see Fig. 10.35). With R_2 and R_1 being made equal, the circuit functions as an *inverting follower* with a gain $K = -1$. The input resistance of such a follower is equal to R_1 .

If we replace R_2 by a capacitor C , the gain of Eq. (10.37) in operator form will be

$$K(s) = -1/(sCR_1)$$

As known, $1/s$ is an integral operator. So the op amp with such a feedback circuit is an *op amp integrator*, in which the relation between the output and input voltage has the form

$$v_{out}(t) \sim \int v_{in}(t) dt$$

If we replace R_1 by C , then, according to Eq. (10.37),

$$K(s) = -sCR_2$$

The operator s is a differential operator, and here the op amp converts to an *op amp differentiator*, in which

$$v_{out}(t) \sim (d/dt) v_{in}(t)$$

It is likewise possible to make up circuits for handling other operations.

10.11. Testing of Integrated Circuits

Integrated circuits, like any other electronic devices, must be fit for work under different operating conditions. This calls for control of ICs during and after manufacture to detect defectives. The conditions of test should resemble as much as possible the conditions at which an IC can or is likely to operate.

Each IC located on the wafer is subjected to a complete checkout prior to scribing. Defective ICs are marked with a drop of paint and rejected after scribing. Sound ICs are then used in assembly operations. After assembling and packaging, the devices are put to various tests such as electrical, design, mechanical, environmental, and, in some cases, radiation tests.

Electrical tests are aimed at measuring most important parameters which enable a device to perform its specified functions. For digital ICs these are logic 1 and logic 0 voltage levels, input currents, delay times, and some other parameters. For analog ICs, these are gain factors, cutoff frequencies, leading edge and trailing edge times, output resistances, and others. For digital LSI circuits, these parameters cannot generally be measured because of a limited number of

external terminals. Therefore, electrical tests on digital LSI circuits are *functional* in character: definite codes are set at the inputs and the reaction to these codes are detected at the outputs. It is practically impossible to locate a concrete defective section or element which causes the malfunction of the entire LSI circuit.

Design tests include a visual check of the IC package and external terminals for dents, scratches on the package, bends, and cracks in the terminals. A test for hermetic sealing of a package holds a special place. In use are two methods for such verification, the method of helium leakage and the method of hot oil. By the first method, a small amount of helium is introduced into the package during its sealing. Helium exhibits the property of superfluidity, that is, ability to issue through minute holes, if any, in the package. Helium leakage is detected by special testers called leak detectors. In the second method, which is simpler and cheaper (but less accurate), an IC is immersed in a vessel filled with hot process oil. If the package is not air-tight, the residual gases get out through holes, and so the bubbles appearing at the surface attest to poor package sealing.

Mechanical tests are divided into vibration and impact types. The first are made on vibration-testing machines. ICs are glued to a vibrating table, which is then set in motion to vibrate at a definite amplitude (to a few millimeters) and a definite frequency (to hundreds of hertz and more). These tests permit detecting unreliable connections between an IC and package terminals. For higher assurance, two test cycles are conducted for vibration resistance with ICs held in two, mutually perpendicular, positions relative to the table. Impact test methods use a pendulum-type impact machine. A heavy pendulum is swung back a certain distance away from an IC fastened on the board and then is allowed to go down. The impact strength is measured in acceleration g-units. The typical values are 20 000 g, but sometimes they range into 10^5 g. The test cycle is usually limited to 2 or 3 impacts.

Environmental tests include test procedures for determining electrical parameters at various conditions of the environment. Of primary importance are the tests for establishing the operating-temperature range. They are run at limiting temperatures (over the range -70°C to $+125^{\circ}\text{C}$ for silicon devices). Heat-cycling tests then follow. The procedure comes to exposure of ICs to low temperatures maintained in a cryostat and then to high temperatures in a thermostat in rapidly changing cycles. Other types of test involve exposure of ICs to increased humidity (98 to 100 %), to see fog (in the atmosphere of dispersed salt solutions), to biological factors (fungi), to low and high pressures (in high-pressure chambers), and to other influences.

A combination of the above tests makes it possible to detect unfit devices and thus supply the customer with the ICs which can in principle perform their functions under specified operating conditions during a definite time after putting them into service. The problem concerning this period of operation and evaluation of the anticipated serviceability of ICs relates to the theory of reliability. Some elements of this theory will be discussed in the section that follows.

10.12. Reliability of Integrated Circuits

It has been pointed out in Ch. 1 that the enhanced reliability of ICs is one of the main factors which ensures the development of microelectronics as a specific field of science and engineering.

Reliability is not an intuitive notion; there are a number of quantitative parameters which in combination characterize the *quality* of integrated circuits. Reliability is the probability of *failure-free performance* of a device for an intended length of time under specified operating conditions.

Failure of an IC (or any other device) is understood to be either a complete inability of the IC to operate or its malfunction such that some of its parameters deviate from the permissible, previously stipulated ratings. Correspondingly, one commonly discriminates between *complete* (catastrophic) and *creeping* (progressive) *failures*. Thus, a shorting between the collector and base may be taken as an example of complete failure, and a threefold decrease in the gain β as an example of creeping failure.

10.12.1. Causes of IC failures. Complete failures essentially result from shortings and bursts in an IC, and creeping failures from progressive changes in conductivity and in other electric and physical parameters of individual parts of the IC.

If shortings and bursts have appeared before the control check, they can be detected and faulty ICs rejected in electrical testing. But if they are still in the incipient stage, the testing procedure will fail to reveal them, and the defects will be **potential** factors of unreliability. Let us enumerate the typical processes causing shortings and bursts.

A short-circuit may appear if, for example, one connection wire comes in contact with another or with the case under mechanical vibrations or impacts; if a circuit portion overheats and melts out; or if a current-conducting substance penetrates into the dielectric through its pores. The last defect often occurs in a thin oxide of the MOS transistor if moisture gets into the IC package. Local overheating is inherent in power transistors and is practically nonexistent in ICs, let alone LSI circuits.

Bursts may result from mechanical influences (vibrations, shocks) and from electrochemical and chemical processes. Vibrations and shocks generally impair electrical contact between connection wires and bonding pads. Electrochemical and chemical processes show up in a variety of ways.

First, *electrochemical corrosion* of metal films and contact connections takes place. For example, in the presence of traces of moisture and hydrochloric acid, aluminum metallization converts to alumina, Al_2O_3 .

Second, chemical processes cause the formation of *intermetallic compounds*. This phenomenon is particularly apparent in contacts between dissimilar metals, for example, between a gold wire and aluminum bonding pad. At one time the malfunction of such contacts led to numerous failures of ICs because of "purple plague", a metal disease responsible for the conversion of metals into a non-conducting powderlike compound of purple color at the boundary between Al and Au. To prevent the formation of such a compound requires performing thermocompression operations under strictly specified temperature conditions.

Third, the process of *electromigration* is to be mentioned. This is the movement of metal (aluminum) atoms into adjacent regions under the action of an electric field and elevated temperature. The thickness of a metal stripe then decreases and an open-circuit appears as a consequence of local overheating and "burning" of the stripe.

Consider now the processes responsible for *progressive* (creeping) failures, or, in essence, for the time drift of IC parameters. Of course, a strictly defined border-line between complete and creeping failures does not exist. It can be said that *a complete failure results from an avalanche-like accumulation of those changes which earlier caused the drift of parameters.*

The processes occurring at the boundary between silicon and a protective oxide play the main part in the origin of creeping failures. They cause the formation of inversion and depletion layers—channels—under the action of ions located in the oxide (see Sec. 3.5). These channels, as noted in Subsec. 3.5.2, have a direct effect on the reverse currents of *pn* junctions and on the value of breakdown voltage. Instability of these two parameters is due to **migration** of ions in the oxide. In turn, ion migration is due to diffusion (particularly at elevated temperature) and drift of ions in electric fields.

Electric fields are inevitably present in the oxide since the latter borders on semiconductor layers and metal stripes which are at different potentials. Both the thickness of the oxide and the width of *pn* junctions average fractions of a micrometer, therefore the field strength reaches 10^4 V/cm and above even at a potential difference of 0.5 to 1 V.

The field direction can be both longitudinal (parallel to the interface) and transverse (perpendicular to the interface). Consequently, the migration of ions in the oxide occurs in both directions, thereby causing changes not only in the conductivity of channels but also in their **length**. For example, if the inversion n channel formed in the base initially terminates short of the collector contact, then, with time, it can form a "bridge" between the base and collector contacts (see Fig. 3.22d). A "bridge" present in a switching circuit will inhibit double injection operation, and so the level V^0 will go up. In an amplifier circuit, a "bridge" will cause a decrease in the input resistance. Either of these defects will grow progressively as longitudinal migration of ions continues and the channel extends further on.

Surface phenomena also affect the values of B and β . Indeed, according to Eq. (4.24), the gain B depends on the values of L_b and L_e . In turns, these values depend on the effective lifetimes of carriers in the base and emitter layers [see Eq. (2.66)]. The effective lifetime is the sum of bulk and surface components [see Eq. (2.41)]. Since ion migration in the oxide changes the surface layer structure and hence the surface recombination rate, the base current gains B and β are liable to time drift.

10.12.2. Methods of reliability evaluation. To date the basic method of IC reliability evaluation has been a statistic method relying on the test of a batch of devices for service life. If n failures have occurred in the batch of N pieces in time t , then the failure probability per unit time is

$$\lambda [1/h] = n/(Nt) \quad (10.43)$$

The quantity λ is known as the *mean failure density*, or *failure rate*.

Knowing the value of λ permits estimating the *probability of failure-free operation* of an IC over a specified period of service:

$$P = e^{-\lambda t} \quad (10.44)$$

From Eq. (10.44) it follows that, whatever small the quantity λ can be, with time the probability of survival approaches zero.

The *mean time to failure*, or *mean life* of a device, is customarily found from the condition $\lambda t = 1$:

$$t_m = 1/\lambda \quad (10.45)$$

Setting $\lambda = 10^{-5}$ 1/h, we get $t_m = 10^5$ h (about 10 years).

Generally speaking, the quantity λ is not constant: it changes with time (Fig. 10.43). The curve $\lambda(t)$ features definite sections: section *I*, representative of the failures due to the rough errors in assembling, surface contamination, etc.; section *II*, where λ is constant, that is, the failures result only from random, uncon-

trollable causes; and section *III*, where λ again rises as a result of inevitable *aging* of the device, that is, as a consequence of chemical and physicochemical processes unavoidable in any real structure. For an IC, the principal factors conducive to the above processes are mutual diffusion of dissimilar metals, cosmic radiation-induced defects, and others.

The mean life given by Eq. (10.45) corresponds to the boundary between sections *II* and *III*. Section *I* is commonly eliminated by the producer who trains the devices before they are delivered to the

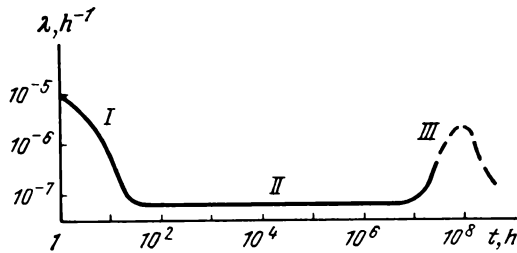


Fig. 10.43. Failure rate versus time

customer. Training is the trial run of devices for a few tens or hundreds of hours in normal operating conditions after they have passed mechanical, electrical, and environmental tests. Training thus helps reveal defectives at the producer's site.

At the present time, the failure rate for ICs and LSI circuits ranges from 10^{-8} to 10^{-9} 1/h. For *reliable* estimation of λ , it is necessary to "wait" for at least 2 or 3 failures in testing. So, at $n = 2$ or 3, as follows from Eq. (10.43), the test procedure for a batch of 10^3 pieces would take tens of years. To cut the test time by using a batch of 10^4 to 10^5 pieces does not prove economical.

In such cases, recourse is made to an accelerated test method based on the *Arrhenius law* which states that the rate of chemical and physicochemical processes, v , is exponentially dependent on temperature:

$$v \sim \exp(-W_a/kT)$$

where W_a is the activation energy of a process. Hence, the mean life at an **elevated** temperature will be substantially shorter than that at the normal temperature:

$$t_{acc} = t_n \exp[-(W_a/k)(T_n^{-1} - T_{acc}^{-1})] \quad (10.46)$$

where subscripts "n" and "acc" relate to normal and elevated absolute temperatures respectively. Accelerated life tests conducted at elevat-

ed temperature **speed up** failure mechanisms, so the device under test fails in a much shorter time. The obtained value of λ_{acc} is used to calculate the value of t_n by expressions (10.46) and (10.45). Testing ICs at a temperature of, say, 250°C can reduce the time it takes to estimate λ by a factor of few hundreds. However, at $\lambda \leq 10^{-9}$ 1/h even such an acceleration proves insufficient. From this it follows that *at the stage of modern microelectronics conventional statistical methods of reliability evaluation become unacceptable*. For the past 5 to 10 years a considerable effort has been directed toward the development of *physical methods* for reliability prediction. These methods involve individual investigations into the structure

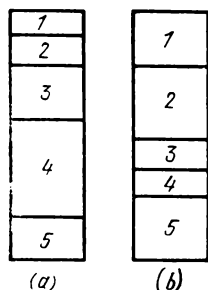


Fig. 10.44. Number of failures due to various causes in simple ICs (a) and large ICs (b)

1—metallization defects; 2—
inaccuracy of diffusion; 3—
defects in chip and oxide; 4—
incorrect application; 5—
other causes

of finished ICs to locate the defects which are *likely to cause* a failure and also investigations on failed ICs to elucidate the *causes* of malfunction and introduce requisite improvements in the fabrication processes.

Unlike statistical test procedures which belong to the category of *destructive* tests, physical test methods are nondestructive and often contactless. The latter include the methods of infra-red imaging, X-ray examination, electro microscopy, and also the methods of measuring the level of excess noise which characterizes the quality of contacts. However, all these methods necessitate complex and costly test equipment and thus cannot as yet be considered conventional. But, considering that statistical methods are now unacceptable, physical methods for LSI reliability prediction will undoubtedly hold the lead in due time.

The failure rate decreases with an increased level of integration because of a higher technological level of LSI circuit manufacture. This brings about *changes in the role of various failure factors* (Fig. 10.44): metallization defects and diffusion inaccuracies (which were relatively insignificant in simple ICs) are in the forefront in LSI circuits. On the contrary, errors due to incorrect application of LSI circuits are in the background (because of a sharp decrease in the number of external connections).

Speaking of the statistical methods of reliability evaluation, we have implied that the test results found from Eq. (10.43) for a *concrete* batch of devices are valid *for other* similar batches. This can only be true if other batches are fabricated according to the same technology as that employed for the tested batch. Hence, an important conclusion follows: *high reliability of ICs is primarily provided by the stability of the manufacturing process*. Any changes

(even progressive) in the manufacturing process may cause a decrease (at least temporary) in the reliability of ICs.

In conclusion, let us stress the fact that the notion creeping defect is conditional. Depending on the type of apparatus, a change in β by 40% can be either unacceptable or acceptable, that is, can or cannot be regarded as a failure. So the ICs considered to be defective (unfit) for use in one device can be employed in another for a still long time.

10.13. Conclusion

From what we have learnt in this chapter, it is now possible to draw the following general conclusions as regards the prospects of microelectronics development in the near future.

An increased scale of integration, which is particularly evident in the last years, is not the end in itself. It only offers the possibility of designing new electronic **devices** (in LSI form) whose functions correspond to those of the former **units** employed in discrete electronics.

One of the unique achievements made in microelectronics in the 1970s was the development of a *microprocessor*. The device comprises an arithmetic-logic and a control unit of a computer fabricated on a single chip (generally, in the CMOS logic, TTL, or I²L form). In combination with an LSI memory, clock pulse generator and data input and output circuits, the microprocessor represents a complete computer disposed on a small printed circuit board. In recent years, an effort made to combine the above units and thus produce a *microcomputer* on a single chip has proved a success. So at present a computer, which formerly belonged to complex systems, has become a kind of circuit component, a constituent of supercomplex systems. It is of interest that by its functional capabilities, a microcomputer is comparable to the first electron tube-based computer of the ENIAC type that came in the late 1940s and occupied an area of $10 \times 13 \text{ m}^2$. This type of computer is much inferior to the microcomputer in speed, reliability, and power dissipation.

The adoption of the microcomputer as a system component marks a new epoch in radioelectronics. But it will take a certain time until the microcomputer comes into wide use (probably in the late 1980s). A further rise in the scale of integration is problematic. In its development, microelectronics will most likely take the course toward functional microelectronics (see below).

The experience gained in the development of radioelectronic apparatus has shown that purely electronic units account for 30 to 40% of the entire units. The remaining 60 to 70% of the units are of the electrical, electromechanical, and mechanical types (light indicators, primary-cell batteries, transformers, relays, electric

motors, electromagnets, connectors, switches, etc.). Microelectronics is obviously far ahead of the related branches as regards its pace of development.

At the present stage, it becomes highly necessary to introduce the methods and means of microelectronics into the related branches to bring about sharp improvements in the size-mass, cost, and reliability indexes of nonelectronic units. This trend which received the name *complex microminiaturization* is gaining ground. New devices have come into being, such as optoelectronic contactless relays, integrated (film) magnetic recording heads, and others.

In the field of electronic units, a purely *quantitative* rise in the scale of integration cannot be infinite. In evidence is a qualitatively new approach which comes to rejecting traditional circuit elements (primarily transistors) in favor of elements using *volume effects* in solids. This approach is likely to lead to higher reliability (survivability).

A classical example of the device relying on these effects is a quartz crystal vibrator which, being a *homogeneous* structure, performs the function of a tank circuit consisting of L , C , and R elements. Other examples may include *Gunn diodes* and *Josephson diodes*, which are homogeneous structures featuring a negative incremental resistance on the current-voltage curve. This feature permits in principle using these diodes as bistable units (that is, in place of multiple-element flip-flop circuits) and also as amplifiers. A number of practical difficulties (in particular, the necessity of maintaining cryogenic temperatures for Josephson diodes) yet makes these approaches problematic. The use of amorphous semiconductors is so far equally problematic (see Subsec. 2.2.4). But the general trend toward the use of volume properties of a solid body should be considered promising.

The above trend goes with the trend toward the use of the volume of liquids, electrolytes. The latter received the name *chemotronics*. On the whole, it can hardly compete with solid-state microelectronics, but some particular approaches are likely to show promise for complex microminiaturization. A general drawback of chemotronic devices is a low speed of response due to the lag in velocity of ions moving in the electrolyte.

A scientific and engineering trend toward the use of volume effects to enable replacing the component structure of ICs conditionally received the name functional microelectronics.

Considering the above discussion, it can be inferred that modern microelectronics is probably a certain complete stage and a new stage will set in sooner or later, which will qualitatively differ from the preceding stage. Obviously, a new stage will rely on the physical, technological, and circuitry principles that underlie modern microelectronics.

REFERENCES

1. Aleksenko, A. G., Fundamentals of microcircuitry. Sovetskoye Radio, 1977 (in Russian).
2. Efimov, I. E., Gorbunov, Yu. I., Kozyr I. Ya., Microelectronics, Moscow, Vysshaya shkola, 1977 (in Russian).
3. Hamilton, D. J., Howard, W. G., Basic Integrated Circuit Engineering, McGraw-Hill, N. J., 1976.
4. Handbook of Semiconductor Electronics. Edited by Hunter, L. P., 3rd edition, McGraw-Hill, 1970.
5. Jang, E. S., Fundamentals of Semiconductor Devices. McGraw-Hill, N. J., 1978.
6. Millmann, J., Microelectronics: Digital and Analog Circuits and Systems. McGraw-Hill, N.J., 1979.
7. Pasyukov, V. V., Chirkin, L. K., Shinkov, A. D., Semiconductor Devices. 2nd edition, Moscow, Vysshaya shkola, 1973 (in Russian).
8. Shalimova, K. V., Physics of Semiconductors. 2nd edition, Moscow, Energiya, 1976 (in Russian).
9. Stepanenko, I. P., Basic Theory of Transistors and Transistor Circuits, 4th edition, Moscow, Energiya, 1977 (in Russian).
10. Streetman, B. G., Solid State Electronic Devices. 2nd edition, Prentice-Hall, 1980.

INDEX

- Accepters, 30
- Admittance,
 - equivalent input, 330
- Amorphous substances, 27
- Amplifiers, 315-323
 - cascode, 354
 - differential, 330-344
 - MOS transistor, 342
 - MOSFET, 325, 326
 - operational, 427-434
 - push-pull, 358-360
- Amplification factor, 160, 317
- Analog circuits, 264, 312
- Anisotropy, 23
- Anodizing, 198
- Artwork, 190
- Assembling of ICs, 201-204
- Asymmetrical triggering, 303

- Band bending, 61, 154
- Band diagrams, 31-39, 79, 80, 93, 107-111
- Base resistance modulation, 91
- Basic constants, 37
- Bipolar transistors, 118-125, 133-138
 - common-base, 119, 127, 131, 146
 - common-emitter, 119, 128, 136, 146, 147
 - diffusion, 117, 120
 - drift, 117, 122
- Bistable units, 301
- Boltzmann constant, 37
- Boltzmann's equilibrium, 42, 43
- Bonding of components, 207, 208
- Boundary,
 - pn junction, 76
- Breakdown in *pn* junctions,
 - avalanche, 93-96
 - thermal, 96
 - tunnel, 93, 96

- Carrier
 - concentration, 36-45, 48, 81, 83, 123
 - diffusion, 67-70
 - distribution, 39, 43, 73, 79, 120-125, 280
 - drift velocity, 48
 - extraction, 73, 85
 - generation, 28, 53
 - injection, 68, 85
 - lifetime, 54-57
 - mobility, 43-47
 - recombination, 33, 51
 - direct, 51-53
 - equilibrium, 52
 - rate of, 52, 53, 57
 - surface, 56
 - trap, 54
 - removal, 103
 - transport, 86
- Carriers,
 - excess, 53, 55, 70, 103
 - free, 36
 - injected, 73
 - intrinsic, 40
 - majority, 30, 43
 - minority, 30
- Capacitance, 153, 165
 - barrier, 97, 148, 164
 - differential, 98
 - diffusion, 97, 99
 - feedback, 242
 - junction (see barrier)
 - overlap, 164
 - parasitic, 293
 - per-unit area, 257
- Capacitors,
 - diffused, 254-257
 - film, 261, 262
 - memory, 304
 - MOS, 258
- Capture cross section, 52
- Charge-coupled devices (CCD), 420, 423
 - buried-channel, 427
 - two-phase, 426
- Collector diffusion isolation (CDI), 271
- Common-mode rejection ratio, 335, 343
- Common-mode signal, 336

- Component attachment, 207
- Concentration gradient, 65, 181
- Conduction, 29
- Conductive pastes, 16
- Conductivity, 43, 48, 49
- Contacts,
 - metal-semiconductor, 106-111
- Contact potential difference, 37, 59, 108
- Continuity equations, 66
- Crystal defects,
 - Frenkel, 24, 25
 - point, 25
 - radiation, 24
 - Schottky, 24, 25
- Crystal lattice, 23
 - dislocation in, 25, 26
- Crystal planes, 23, 24, 27
- Crystal surface, 26
- Current, 65
 - diffusion, 65, 71
 - drift, 43
 - forward, 76
 - input offset, 340, 344
 - reverse, 76, 105
 - thermal, 87, 88
 - thermally generated, 92
- Current gain, 127-132, 140, 145, 147, 148
 - transistor, 127, 224
- Current reflectors (mirrors), 370-372
- Current regulators, 366
- Cutoff frequency, 145, 224
 - current gain, 148
 - of transconductance, 166
- Current-voltage (I-V) characteristics,
 - pn junction, 85-96
 - switch, 280, 288
- Czochralski technique
 - of crystal pulling, 174

- Darlington pair, 313
- Debye length, 58
- Delay time, 143, 274
- Dember effect, 68
- Depletion layer, 58-62, 79, 152
- Depletion mode, 63, 64
- Deposition
 - electrolytic, 199
 - thick film, 209, 210
 - thin film, 194-199
- Detector,
 - voltage level, 312
- Differential signal, 332
- Diffusants, 180
- Diffusion, 215, 216
 - carrier, 67-70
 - double, 246
 - selective, 179
 - total, 179
- Diffusion
 - coefficient, 183
 - constant, carrier, 57, 66
 - depth, 183
 - equations, 67, 69
 - furnace, 180
 - length, 72
 - theory, 181
 - time, 100, 144
- Digital circuits, 264
- Diodes (see also *pn junctions*),
 - integrated, 237
 - point-contact, 106
 - reference, 96, 238
 - Schottky, 109
 - semiconductor, 77
- Diode-transistor logic (DTL), 381
- Direct-coupled transistor logic (DCTL), 376
- Donors, 29
- Doping, 178
 - depth of, 74
- Drift
 - compensation, amplifier, 338, 434
 - of dc components, amplifier, 319
 - velocity, carrier, 48
- Dynamic MOS (DMOS) logic, 394

- Early effect, 132
- Ebers-Moll model, 134
- Ebers-Moll equations, 135
- Electronics, definition of, 9
- Electronic circuits (see also ICs),
 - analog, 264
 - classification of, 264
 - digital, 264
 - inverting, 265
 - noninverting, 265
 - switching, 271
- Emitter-coupled transistor logic (ECL), 379
- Emitter follower, 344, 345, 352, 365
- Encapsulation, 203
- Energy levels, 30, 36, 38
 - density of, 39, 40
 - distribution of, 39
 - surface, 32
 - trap, 36
- Enhancement mode, 63
- Enriched layer, 62, 115, 152
- Epitaxial passivated IC (EPIC) process, 218
- Epitaxy, 174-176
 - liquid phase, 176
 - vapor phase, 175, 176

- Equivalent circuits,
 - amplifier, 316
 - capacitor, 256
 - resistor, 252, 253
 - switch, 277
 - transistor, 133, 271, 272
- Error function, 72
- Etching, 185, 208
 - anisotropic, 187
 - electrolytic, 186
 - ionic, 186
- Evaporation,
 - thermal (vacuum), 194
- Exposure time, 184
-
- Fall time, 274
- Fan-in, 399
- Fan-out, 399
- Fermi level, 39, 41, 44, 64
- Fick's laws, 181
- Field effect
 - in semiconductors, 58, 60-63
- Field-effect transistors (FETs), 116, 150
 - CMOS, 152, 243
 - junction, 168-172, 239
 - MNOS, 248
 - MOS, 14, 15, 151-160, 243
 - built-in channel, 153
 - self-aligned poly-Si gate, 245
- Field-form factor, 74
- Flip-chip method, 207
- Flip-flops, 301
 - D, 405, 406
 - JK, 405
 - T, 306, 403, 404
 - RS, 400
 - RST, 402
- Frequency characteristics, 142
 - amplifier, 324
 - bipolar transistor, 145, 146
 - JFET, 172
 - MOSFET, 163-166
-
- Gain,
 - current, 127-132, 140, 145, 147, 148, 224
 - common-mode, 335, 343
 - difference-component, 334
 - voltage, 317, 334, 345
- Gates, logic (see logic elements)
-
- Hybrid ICs, 12-16
 - compatible, 13
 - large-scale, 419
 - thick film, 16, 209
 - thin film, 17, 204, 207
-
- Impact ionization
 - coefficient, 95
- Impurities, 25
 - interstitial, 25
 - substitutional, 25, 29
- Impurity
 - concentration, 45-47, 114, 115
 - distribution, 82
 - distribution, 181-184, 221, 222
 - solid solubility, 180
- Inductors, film, 263
- Injection efficiency, 128, 129
 - time constant of, 148
- Injection level, 69
- Injection operation, transistor,
 - double, 124, 125
- Integrated circuits (ICs), 12-21, 374
 - assembling of, 201-204
 - classification of, 12
 - complementary, 120
 - definition of, 9
 - hybrid, 13-17, 207, 209, 419, 440, 442
 - I²L, 389
 - large-scale, 16, 413, 419
 - manufacture of, 11, 204, 209
 - medium-scale, 16
 - MOS, 14, 15
 - packaging of, 203
 - semiconductor, 12, 14-16, 194
 - small scale, 16, 413
 - thick-film, 12, 16, 209
 - thin-film, 12, 17, 194, 207
 - very large-scale, 16, 414
- Integrated diodes, 237
- Integrated elements, 212
 - definition of, 9
 - film, 259
 - isolation of, 213-219
- Integrated injection logic (I²L), 385
 - circuits, geometry of, 389
- Interconnection pattern, 14, 200
- Interface, 112-115
- Inverse operation, transistor, 125
- Ion implantation, 184, 185
- Isolation methods, 214-219
 - air, 219
 - CDI, 217, 231
 - diffusion, 216-218
 - isoplanar, 219, 220
 - pn junction, 215, 216
 - SOSIC, 219
 - V-groove, 221

Junction (see *pn* junction)

Junction FETs (see FETs)

Laplace transform, 71, 144

Large-scale integration (LST), 16, 413, 419

characteristics, of, 414

Level shifters, 353

Logic,

CMOS, 292

diode-transistor, 381

direct-coupled transistor, 376

DMOS, 394

emitter-coupled transistor, 379

integrated junction, 385

resistor-capacitor transistor, 379

resistor-transistor, 378

transistor-transistor, 383-385

Logic

elements, 374

bipolar, 374-385

MOS, 391, 396-399

functions, 374

swing, 376

symbols, 375

Masking, 188

Masks, 14, 188, 189, 20, 4209

photoresist, 200, 208

Maxwell-Boltzmann

distribution function, 39

Medium-scale integration (MSI), 16

Memories,

random access, 407, 411

programmable, 413

read-only, 407, 411

Metal-nitride-oxide silicon (MOS)

transistor, 248

Metal-oxide-semiconductor (MOS)

transistor (see FETs)

Metallization, 14, 199, 417

Microelectronics,

definition of, 9

Miller effect, 242, 329

Miller indexes, 23

Monolithic ICs (see semiconductor ICs)

MOS logic, 390

Noise immunity, 297, 399

switch, 297

Nyquist formula, 172

Packaging, 203

Pattern,

interconnection, 200

Pattern

alignment, 190, 194

generator, 193

Photomasks, 188, 189-192

Photomasking, 189, 205

Photomask alignment, 190

Photolithography, 188

electron beam, 193

projection, 193

resolution of, 192

Photoresist, 208

PN junctions, 77, 78

abrupt, 77

asymmetric, 77

breakdown, 93

equilibrium, 80

graded, 77

isolation, 215

nonequilibrium, 83

one-sided, 77

planar, 114

rectifying, 106

symmetric, 77

Poisson equation, 60, 61

Polycrystal, 26

Potential,

chemical, 41

electrochemical, 41

electrostatic, 39, 60

surface, 59

thermal, 38

Potential barrier, 80, 81

Pulse generator, 311

Pulse shaper, 312

Push-pull circuits, 357-360

Random access memories (RAMs), 407

Read-only memories (ROMs), 407

Recovery time, 105

Relaxation time, 36

Reliability of ICs, 442

Resistance,

base, 91, 141

channel, 159

collector, 141, 149

emitter, 141

incremental, 91

input, 318, 337, 346

lateral, 224

leakage, 257

output, 319, 347

sheet, 223

thermal, 97

Resistors,

ballast, 364

film, 260

semiconductor,

- diffused, 249, 250
- ion-doped, 251
- pinch, 250
- Resistor-capacitor-transistor logic (RCL), 379
- Resistor-transistor logic (RTL), 379
- Rise time, 274
- Scale of integration, 16, 415
- Schmitt trigger, 308
- Schottky barrier, 108
- Schottky barrier transistor, 166, 232
- Schottky diode, 109
- Screen printing, 209
- Scribing, 201
- Semiconductors, 23, 34
 - band structure of, 30
 - compensated, 36
 - degenerate, 32, 109
 - extrinsic, 32, 35, 50, 62, 63
 - inhomogeneous, 48, 73
 - intrinsic, 28, 35, 60-62
 - n*-type, 29
 - p*-type, 30
 - wide-gap, 45
- Semiconductor ICs
 - classes of, 12-16
- Semimetals, 32
- Sensitivity threshold, 297
- Silicon, 22
 - breakdown in, 95
 - conductivity of, 49, 50
- Silicon-on-sapphire IC (SOSIC) process, 219, 243
- Small-scale integration, 16, 413
- Small-signal models,
 - amplifier, 317, 322
 - emitter follower, 415
 - transistor, 140, 164, 171
- Solid solubility, impurity, 180
- Sputtering,
 - cathode, 194-196
 - ion-plasma, 194, 197
- Static characteristics,
 - bipolar transistor, 133-138
 - JFET, 168-170
 - MOS, 155
 - switch, 267
- Stencil screen, 16, 17, 209
- Storage time, 103, 278
- Surface states, 32, 37
- Switches,
 - bipolar transistor, 266-277
 - CMOS transistor, 288, 291, 296
 - current, 284,
 - MOS transistor, 288
 - dynamic load, 289, 290
 - resistive load, 288
- Switching circuits, 271
- Symmetrical triggering, 303
- Testing of ICs, 440
- Thermal oxidation, 176
 - types of, 177
- Thermionic emission, 33
- Thermocompression bonding, 202, 203
- Time constants, 149, 166, 177
- Transconductance, 161, 166, 171
 - critical, 163
 - specific, 157
- Transfer characteristics,
 - amplifier, 359
 - electronic circuit, 265
 - Schmitt trigger, 312
 - transistor, 157
- Transients in
 - amplifiers, 321, 324, 329, 341
 - diodes, 101
 - emitter followers, 348-351
 - flip-flops, 306
 - pn* junctions, 97-106
 - switches, 274, 287, 292-296
 - transistors, 142-149, 163-166
- Transistors, 11, 14, 15, 239, 243-247
 - ball-lead, 207
 - beam-lead, 208
 - bipolar, 116, 118-147
 - composite, 313
 - field-effect, 116, 150
 - fabrication of, 239, 243-247
 - npn*, 221-223, 227
 - multicollector, 231, 387
 - multiemitter, 229
 - parameters of, 223, 224
 - parasitic, 225
 - Schottky barrier, 232
 - superbeta, 233
 - pnp*, 234
 - composite, 314
 - parasitic, 234
 - topology of, 235
 - unipolar (see field-effect)
- Transistor switch (see switches)
- Transistor-transistor logic (TTL), 383
- Transit time, 144
- Transition time, 103
- Transport factor, 128
- Trigger circuit, 308, 312
- Trigger levels, 309, 312
- Tunnel effect, 59
- Undercoat, 195, 199, 205
- Unipolar transistors (see FETs),

- Very large-scale integration (VLSI),
16, 414
- V-groove isolation, 221
- Voltage,
 - breakdown, 95, 224
 - cut-off, 153
 - Dember, 68
 - flat-band, 59
 - forward, 83, 102
 - input offset, 328
 - residual, 288
 - reverse, 83
 - threshold, 152, 153, 157, 247
- Voltage comparator, 437
- Voltage gain,
 - amplifier, 317, 334
 - emitter follower, 345
- Voltage regulators, 361, 437
 - diode, 362
 - transistor, 364
- Work function, 33, 108

TO THE READER

Mir Publishers welcome your comments on the content, translation and design of the book.

We would also be pleased to receive any suggestions you care to make about our future publications.

Our address is:

USSR, 129820, Moscow, I-110, GSP, Pervy
Rizhsky Pereulok, 2, Mir Publishers.

